UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI

Programa de Pós-Graduação em Educação

Leila Maria Silva

ENEM NAS REDES SOCIAIS: Mineração de Textos e *Clusterização*

Diamantina
2017

Leila Maria Silva

ENEM NAS REDES SOCIAIS: Mineração de Textos e *Clusterização*

Trabalho apresentado ao Programa de Pós Graduação em Educação da Universidade Federal dos Vales do Jequitinhonha e Mucuri, para obtenção do título de Mestre em Educação.

Orientador: Prof. Dr. Marcus Vinicius Carvalho Guelpeli

Diamantina
2017

Ficha Catalográfica - Sistema de Bibliotecas/UFVJM
Bibliotecária: Jullyele Hubner Costa CRB-6/2972

S586e Silva, Leila Maria.
ENEM nas redes sociais: mineração de textos e clusterização / Leila Maria Silva, 2018.
90 p.

Orientador: Marcus Vinicius Carvalho Guelpeli

Dissertação (Mestrado Profissional - Programa de Pós-Graduação em Educação) –Universidade Federal dos Vales do Jequitinhonha e Mucuri. Teófilo Otoni, 2018.

1. Mineração de textos. 2. Twitter. 3. ENEM. 4. Redes Sociais. 5. Cassiopeia. I. Guelpeli, Marcus Vinicius Carvalho. II. Universidade Federal dos Vales do Jequitinhonha e Mucuri. III. Título.

CDD 371

Elaborada com os dados fornecidos pela autora.

LEILA MARIA SILVA

ENEM NAS REDES SOCIAIS: MINERAÇÃO DE TEXTOS E *CLUSTERIZAÇÃO*

Dissertação apresentada ao PROGRAMA DE PÓS-GRADUAÇÃO EM EDUCAÇÃO - STRICTO SENSU, nível de MESTRADO como parte dos requisitos para obtenção do título de MAGISTER SCIENTIAE EM EDUCAÇÃO

Orientador : Prof. Dr. Marcus Vinícius Carvalho Guelpeli

Data da aprovação : 18/12/2017

Prof.Dr. MARCUS VINÍCIUS CARVALHO GUELPELI - UFVJM

Prof.Dr. ALEXANDRE RAMOS FONSECA - UFVJM

Prof.ª Dr.ª GERUZA DE FÁTIMA TOMÉ SABINO - UFVJM

Prof.ª Dr.ª MARIA LUCIA BENTO VILLELA - UFVJM

DIAMANTINA

RESUMO

A *internet* é hoje a maior fonte de informação eletrônica existente. Cresce a cada dia o número de usuários da internet, e consequentemente o uso das redes sociais *online*. São muitas as informações novas que ficam embutidas nas bases de dados textuais. Por causa da sua natureza dinâmica, ou seja, milhões de páginas surgem e desaparecem todos os dias, a tarefa de encontrar informações relevantes nessas bases de dados se torna muito difícil. As técnicas de mineração de textos para a descoberta de informações na *web* surgiram da necessidade de sanar este problema. O presente trabalho versa sobre a aplicação de métodos de mineração de textos com clusterização na grande quantidade de mensagens sobre o Exame Nacional do Ensino Médio no ano de 2016 provenientes da rede social *Twitter*. O foco deste estudo está na obtenção de grupos de textos, a fim de possibilitar uma visualização resumida e sintetizada dos assuntos mais comentados pelos usuários. Para manipulação dessas bases textuais, o Modelo Cassiopeia foi utilizado empregando seu algoritmo de agrupamento textual que tem como principal finalidade gerar agrupamentos, ou seja, *clusters* (grupos) de documentos textuais que apresentam algum tipo de similaridade. O Modelo Cassiopeia apresenta um limite de processamento com a quantidade máxima de 700 *tweets*. Os *tweets* passam primeiramente pela fase de limpeza dos textos no pré-processamento, logo após, a utilização do algoritmo no processamento e por fim, as análises dos resultados no pós-processamento. Os resultados obtidos neste trabalho mostram valores coesos quanto à similaridade dos documentos dentro de um *cluster* e entre os *clusters*, avaliados por medidas de agrupamento textual, proposto pelo Modelo Cassiopeia. Isso demonstra a aplicabilidade dessa proposta para a visualização sintetizada das informações mais significativas de um determinado tema, muitas vezes permitindo que ações sejam antecipadas e impactos sobre a população afetada sejam reduzidos.

Palavras-chave: Mineração de textos, *Twitter*, ENEM, *Cluster*ização, Redes Sociais, Cassiopeia

ABSTRACT

The Internet is today the largest source of existing electronic information. The number of Internet users is increasing daily, and consequently the use of online networks online. There are many new information that is embedded in textual databases. Because of its dynamic nature-that is, millions of pages and other numbers-a task of finding relevant information in those databases becomes very difficult. The techniques of text mining for a discovery of information on the web came from the need to heal this problem. The present work is about an application of methods of text mining with clustering in the large amount of messages on the National High School Exams in the year 2016 issu social network Twitter. The focus of this study is on obtaining groups of texts in order to enable a summary and synthesized publication of the appropriate comments of the users. For manipulation of textual bases, the Cassiopeia Model was used by using its textual grouping algorithm that has as main purpose to generate clusters, that is, clusters of textual documents and executed some kind of similarity. The Cassiopeia Model has a processing limit with a maximum of 700 tweets. The tweets first pass through the phase of cleaning the texts without preprocessing, afterwards, a use of the algorithm without processing and, finally, as analysis of the results without post-processing. The results obtained in this work are more closely related to the similarity of the documents within the cluster and between the clusters, through the measurements of textual grouping, proposed by the Cassiopeia Model. This demonstrates an application for an uninformed publication of the most important information on a given topic, often allowing actions to be anticipated and impacts on an affected population to be reduced.

Keywords: Text mining, *Twitter*, ENEM, *Cluster*ing, Social **N**etworks, Cassiopeia

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ENEM - Exame Nacional do Ensino Médio

FIES - Fundo de Financiamento Estudantil

HTML - *HyperText Markup Language*

INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

KDT - *Knowledge Discovery in Texts*

LABIC - Laboratório de estudos sobre Imagem e Cibercultura

MEC - Ministério da Educação

MTPLNAM - Mineração de Texto, Processamento de Linguagem Natural e Aprendizado de Máquina

PROUNI - Programa Universidade Para Todos

SISU - Sistema de Seleção Unificada

UFES - Universidade Federal do Espírito Santo

UFRJ – Universidade Federal do Rio de Janeiro

UFVJM - Universidade Federal dos vales do Jequitinhonha e Mucuri

SUMÁRIO

1 INTRODUÇÃO

As redes sociais ganharam grande destaque nos últimos anos. A comunicação nas redes sociais surgiu da necessidade que os seres humanos têm, em compartilhar com o outro, suas preferências sobre diversos assuntos como futebol, culinária, músicas, etc, criando assim, laços de afinidades entre eles (WIVES, 2004).

Com o constante crescimento do uso dos meios tecnológicos, existe uma grande quantidade de informações sendo armazenadas e processadas por meios computacionais. Em consequência disso, estas bases textuais passam a conter informações ricas sobre vários dos procedimentos de organizações e, principalmente, sobre seus usuários. Essas informações representam um ativo importante para a tomada de decisão.

Porém as grandes bases textuais por si só, são apenas dados isolados para os computadores. Os mesmos a tratam apenas como uma sequência de caracteres. Faz-se necessária a aplicação de técnicas para estruturar esses dados textuais visando facilitar o conhecimento dos respectivos dados e transformá-los em uma informação para extração de conhecimento.

Dentro do contexto de aprendizagem, pode-se observar uma crescente adoção das redes sociais como recurso de apoio à construção do conhecimento. Elas provêm mecanismos para o compartilhamento de ideias e discussão de temas (LIMA; MOURA, 2014). A análise desses dados textuais oriundos das redes sociais pode auxiliar no levantamento de informações implícitas. É possível examinar as mensagens postadas por meio de técnicas de mineração de textos para posterior análise e reconhecimento de padrões.

A rede social *twitter* tem se destacado pela instantaneidade na divulgação das informações entre pessoas com interesses comuns. É um meio de comunicação aberta, rápido que permite a colaboração e o compartilhamento de ideias em tempo real de informações entre indivíduos e grupos. O *twitter* tem um tesouro de informações sobre comportamentos dos usuários e de tendências em níveis local e global. Devido a essa grande quantidade de informações e a constante presença dos internautas nas redes sociais para expressar suas opiniões sobre produtos, marcas, costumes e preferências, despertou um grande interesse por parte de empresas e pessoas em analisar essas informações que podem ser úteis em pesquisas e estratégias de *marketing* (LIMA; MOURA, 2014).

A partir desse cenário, surgiram ferramentas capazes de analisar as informações contidas nas grandes bases de dados geradas nas redes sociais. Entre essas técnicas está à Mineração de Textos, utilizada na identificação de informações relevantes em grandes volumes de textos de forma clara. Seu emprego consiste em pesquisar, organizar e expor resultados que facilitem o processo de planejamento estratégico a partir dessas informações.

Segundo Aranha e Passos (2006) Mineração de Textos, também chamado de Mineração de Dados Textuais (*Text Data Mining*) é um campo novo e multidisciplinar que mescla conhecimentos das áreas de Informática, Estatística, Linguística e Ciência Cognitiva.

O processo de Mineração de Textos consiste em extrair regularidades, padrões ou tendências de grandes volumes de textos. Inspirado pelo *Data Mining* ou Mineração de Dados, que procura descobrir padrões emergentes de banco de dados estruturados, a mineração de textos pretende extrair conhecimentos úteis de dados não estruturados ou semi-estruturados que são obtidos da *web* (LUCHESI; BERTOLA; ARAÚJO, 2006).

O Exame Nacional do Ensino Médio - ENEM é o maior exame do Brasil. Nos últimos anos, inúmeras funções foram atribuídas ao ENEM, merecendo destacar que, desde 2009, o exame tornou-se uma das principais portas de entrada para o ensino superior no Brasil, atraindo a atenção da sociedade e gerando grande interesse público pela divulgação de dados do exame. Desde que o exame se tornou o principal acesso às instituições de ensino, o ENEM tem passado por alguns problemas que causam temor e descrença entre os estudantes. O primeiro problema foi o furto das provas ocorrido em 2009. Posteriormente, houve uma sucessão de erros como gabarito divulgado com erros, vazamento do conteúdo da prova, candidatos com local de prova alterado, e confusão em aplicação do exame em vários pontos do país. Sendo o ENEM um exame de grande relevância, é interessante obter meios de conhecer melhor todo o processo relacionado a ele com o intuito de amenizar os problemas.

Outra observação é que avaliações como o ENEM estruturam-se com a intencionalidade de conhecer a realidade da educação brasileira, para definir estratégias que favoreçam os objetivos educacionais. Estes modelos de avaliação geram um gigantesco volume de informações, e estas acabam muitas vezes restritas aos organizadores dos programas, ou, quando muito, aos gestores das escolas (VIANNA, 2003).

A proposta deste trabalho é aplicar métodos de mineração de textos com clusterização na grande quantidade de mensagens sobre ENEM 2016 provenientes de redes sociais,

no caso o *Twitter*. O foco deste estudo está na obtenção de grupos de textos, a fim de possibilitar uma visualização resumida e sintetizada dos assuntos mais comentados pelos usuários.

1.1. Motivação

Em análise da literatura levantada, há diversos trabalhos correlatos na área de Mineração de Textos com utilização de técnicas de Agrupamentos. Todos esses trabalhos têm como principal finalidade a descoberta de conhecimento útil através de grandes bases de dados textuais existentes na *internet* e em outras fontes. Eles apresentaram resultados interessantes e eficazes fazendo com que a área de Mineração de Textos tornasse um objeto de estudo bastante útil.

As redes sociais têm ampliado as possibilidades de obtenção de informações proporcionando uma disponibilização em massa de conteúdos na web, podendo estes servir de apoio para a tomada de decisão em várias organizações.

Todavia, buscar e filtrar esses dados produzidos para extrair informações opinativas relevantes a fim de auxiliar uma organização ou usuários em tomadas de decisão não é uma tarefa trivial devido a questões como, por exemplo, o grande volume de dados envolvidos.

Acredita-se que existe um grande potencial de pesquisa no descobrimento de informações úteis dos *tweets*[1] coletados na rede social sobre o tema ENEM, tendo em vista que é um evento de grande relevância no Brasil e interesse social, e também cercado de polêmicas devido a vários problemas que vêm acontecendo desde 2009, quando o ENEM passou a substituir o vestibular e servir como prova de acesso à graduação em diversas universidades federais em todo o Brasil. A importância de estudar essa avaliação dentro do cenário educacional brasileiro impõe-se pela necessidade de instrumentos mais precisos e o adequado conhecimento do processo avaliativo do ENEM aliado ao compromisso com a qualidade da educação brasileira.

[1] ***Tweet*** é o nome utilizado para designar as publicações feitas na rede social do *Twitter*.

1.2. Problema

É possível coletar dados textuais sobre o ENEM 2016 da rede social *Twitter* e aplicar a técnica computacional de Mineração de Texto com *Clusterização* para gerar *clusters* de *tweets* e analisá-los para gerar conhecimento significativo?

1.3. Hipótese

A recuperação de *tweets* da Rede Social *Twitter* com a aplicação da Técnica Mineração de Texto complementarmente usando *Clusterização* através do Modelo Cassiopeia é possível gerar conhecimento explícito agrupado sobre o ENEM 2016 e apresentar resultados satisfatórios das métricas internas (coesão, acoplamento e coeficiente silhouette) sobre os *clusters* gerados.

1.4. Contribuição

A aplicação da técnica mineração de textos em *tweets* sobre o tema ENEM 2016 obtidos da rede social *Twitter*, oferece *clusters* de textos a serem explorados, no intuito de conhecer os assuntos mais comentados pelos usuários e oferecer subsídios para a criação e/ou fortalecimento de estratégias de melhoria do processo relacionado ao ENEM.

1.5 Objetivos

1.5.1 Objetivo Geral

A presente pesquisa tem como objetivo principal analisar a opinião dos usuários do *Twitter* sobre o Exame Nacional do Ensino Médio (ENEM) 2016 através da técnica de Mineração de Textos usando *Clusterização* para extrair conhecimento inerente aos textos analisados.

1.5.2 Objetivos Específicos

- Compreender o conhecimento do evento ENEM 2016 através de mensagens postadas pelos usuários do *Twitter*;
- Aplicar a *Clusterização* sobre os *tweets* utilizando o Modelo Cassiopeia para gerar os aglomerados de *tweets*;
- Validar os resultados obtidos na fase de análise através dos agrupamentos de *tweets* criados sobre o ENEM no ano de 2016.
- Mostrar como utilizar o *Twitter* para analisar as opiniões dos usuários baseado em suas postagens na rede social.

1.6. Estrutura do trabalho

O Capítulo 2 apresenta os trabalhos relacionados.

No Capítulo 3: Fundamentação Teórica serão descritos os principais conceitos que fundamentam este trabalho. Serão apresentados conceitos sobre Descoberta de Conhecimento em Bases de Textos (*Knowledge Discovery in Texts* - KDT), *Clusterização, Twitter,* ferramentas utilizadas como o Modelo Cassiopeia e por fim *corpus*.

No Capítulo 4: Metodologia será apresentado o processo de criação do *corpus* sobre o ENEM 2016, pré-processamento, processamento e estatísticas do *corpus*.

O Capítulo 5: Resultados mostrará os resultados através de gráficos com as métricas obtidas no experimento.

No Capítulo 6: Conclusões serão discutidas as limitações, as contribuições e os trabalhos futuros.

2 TRABALHOS RELACIONADOS

A seguir, serão apresentados alguns trabalhos que se relacionam ao contexto deste trabalho. Existem na literatura científica diversos trabalhos que utilizaram a técnica de Mineração de Textos para diferentes aplicações em diferentes áreas de conhecimento.

Rodrigues (2016), em seu trabalho realizou o processo de mineração de textos em dados do *Twitter* com ênfase na dinamicidade dos assuntos abordados em redes sociais em um determinado período de tempo. Primeiramente, Rodrigues (2016) coletou *tweets* sobre assuntos relacionados com os protestos que ocorreram no Brasil em 2014, como a Parada Gay. Em seguida, realizou a etapa de pré-processamento dos *tweets* e executou o algoritmo de clusterização de dados DBSCAN. Rodrigues (2016) apresentou os resultados do seu trabalho utilizando nuvens de palavras.

Em 2012, Gomide desenvolveu um estudo sobre análise de conteúdo de mensagens divulgadas em redes sociais para detectar eventos e prever eventos da vida real. O estudo mostrou que as mensagens relacionadas à epidemia de dengue e enchentes foram utilizadas para disseminar informação de fontes de opiniões e experiências. Este estudo verificou que as mensagens podem ser usadas para análise de conteúdo em tempo real e de fontes de opiniões e experiências, fornecendo assim um auxílio às autoridades de saúde pública a captarem com maior facilidade as preocupações do público.

Santos (2016), também realizou mineração de texto em seu trabalho que abordou o evento conhecido como “Black Friday” no *Twitter*, ocorrido no dia 27 de novembro de 2015, entretanto, também fez a análise de sentimento das opiniões das pessoas sobre esse evento quanto a polaridade das mesmas, isto é, se as opiniões são positivas ou negativas. Para isso, serão utilizadas técnicas de mineração de texto e análise de sentimentos.

A proposta deste trabalho se assemelha aos estudos apresentados em Rodrigues (2016), Gomide (2012) e Santos (2016) levando em consideração que analizaram mensagens coletadas de redes sociais. Entretanto, são trabalhos que se distinguem em tema, época no processo de coleta e técnicas empregadas.

3 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os principais conceitos que fundamentam este trabalho. Serão abordados os conceitos introdutórios que norteiam uma aplicação de Mineração de Textos, as fases que compõem o processo de Mineração de Textos com *Clusterização*. Também apresenta a contextualização do trabalho, as ferramentas e técnicas utilizadas para o processo de Mineração de Textos.

3.1 Descoberta de Conhecimento em Bases de Textos

A *web* provê maneiras convenientes para as pessoas se comunicarem, expressarem opiniões sobre qualquer assunto e conversar com outras pessoas de qualquer lugar do mundo por meio das redes sociais *online*. Essas redes fazem parte do dia a dia de milhões de pessoas e proporcionam um meio de comunicação que é mundialmente difundido. Cada vez mais pessoas utilizam as redes sociais online para interagir, opinar e compartilhar conteúdos sobre os mais diversos tópicos, que variam desde diversão, clima, trabalho, trânsito, até sua própria condição de saúde (GOMIDE, 2012).

Em plataformas como o *twitter*, os usuários tendem a expressar livremente através das hashtags, o que cria um meio ideal de se capturar as opiniões comuns sobre diversos tópicos. Estes dados textuais, na maioria das vezes, incluem informações valiosas, como exemplo: tendências, anomalias e padrões de comportamento que podem ser usados para auxiliar nas tomadas de decisões (LIMA; MOURA, 2014). Esta informação, no entanto, acaba por se tornando muito extensa inviabilizando que a análise na íntegra de todas as opiniões expressadas. Neste cenário o uso de ferramentas automáticas capazes de extrair o sentimento geral contido nos dados se torna muito atrativo.

Visando transformar estes dados em conhecimento, surge o processo chamado de Descoberta de Conhecimento em Textos, Mineração de textos, *Knowledge Discovery in Texts* (KDT), que Morais (2007) define como sendo “o processo, não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos textos armazenados em um banco de dados”.

Inspirado pelo *Data Mining* ou Mineração de Dados, que procura descobrir padrões emergentes de banco de dados estruturados, a Mineração de Textos, também conhecida como *Text Mining*, pretende extrair conhecimentos úteis de dados não estruturados.

A Mineração de Textos aplica as mesmas funções analíticas da Mineração de Dados (GOMES, 2013), porém para dados textuais não estruturados. Os dados textuais englobam uma vasta e rica fonte de informação, mesmo em um formato que seja difícil de extrair de maneira automatizada.

A KDT vem solucionar grande parte dos problemas relacionados à busca, recuperação e análise de informações. A sobrecarga de informação é um dos maiores problemas enfrentados pelos usuários da *Internet* (WIVES, 2004).

3.2 Técnicas de Mineração de Textos

Para extrair o conhecimento de uma base de textos, podem-se utilizar diversas técnicas de Mineração de Textos. A seguir são apresentadas características das tarefas de classificação, sumarização, associação e *Cluster*ização.

A Técnica de Classificação consiste em examinar as características de um objeto e enquadrá-las em conjuntos pré-definidos. Algoritmos de classificação indicam a correlação de produtos e, a partir do conhecimento do fato, uma ação é tomada.

Este algoritmo analisa todos os exemplos de documentos, aprende as regras e as armazena em uma base de conhecimento. Os documentos a serem classificados passam por um categorizador, o qual, baseado em regras previamente inseridas na base de conhecimento, estabelece a qual classe pertence cada documento (CORRÊA, 2003).

A técnica de Sumarização envolve métodos para selecionar as informações mais importantes do texto, tornando a descrição mais compacta para um subconjunto de dados, mas mantendo a informação a mesma. É bastante utilizada na descoberta de conhecimento em textos, visando identificar palavras ou frases mais importantes do documento ou conjunto de documentos que sumarizam o conceito dos documentos. É útil para reduzir a quantidade de material em um documento, embora mantenha a mesma informação (CORRÊA, 2003).

As regras de Associação possibilitam encontrar regras em um conjunto de dados, do tipo X=>Y, ou seja, transações do banco de dados que contêm X tendem a conter Y. Parâmetros de suporte e confiança devem ser fornecidos para que as regras que satisfaçam esses parâmetros sejam encontradas. Esta técnica é bastante utilizada em mineração de textos, com o objetivo de descobrir as associações existentes entre termos e categorias de documentos (CORRÊA, 2003).

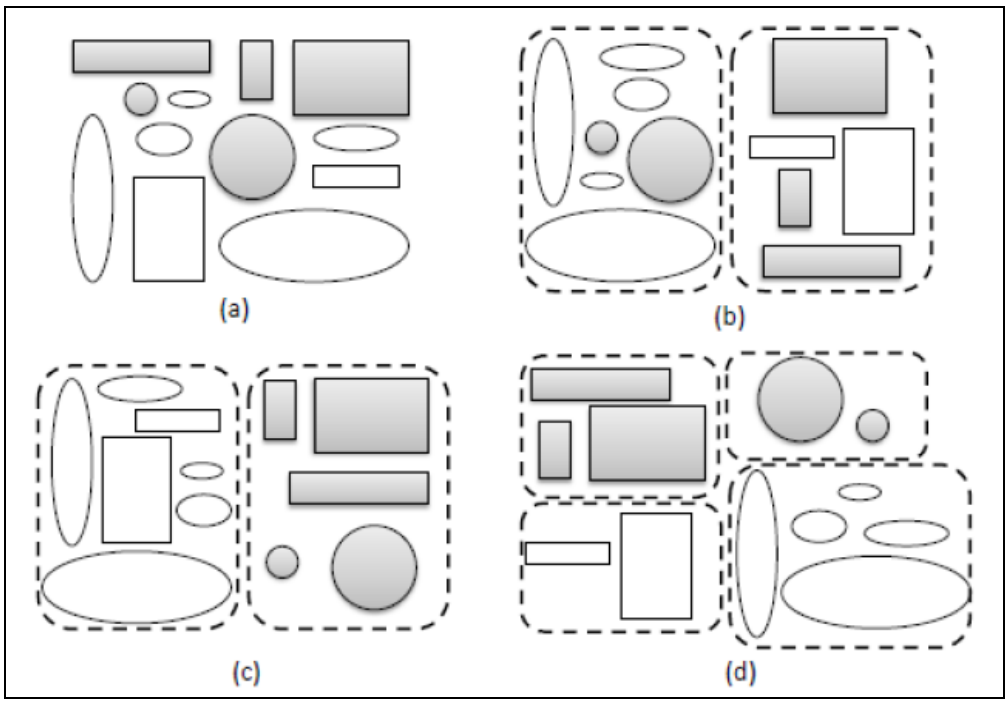
A Técnica de *Clusterização* realiza um processo capaz de identificar documentos similares e alocá-los em grupos, denominados *Clusters*. Um *Cluster* é um conjunto de objetos similares entre si e, ao mesmo tempo, diferentes de objetos presentes em outros conjuntos.

Seu objetivo é, portanto, ter um maior conhecimento sobre esses dados e suas relações. Assim, o processo de agrupamento consegue agrupar uma coleção de padrões desconhecidos (não classificados) em conglomerados (*Clusters*) que possuam algum significado para o usuário (WIVES, 2004).

Na técnica de agrupamento ou *Clusterização*, documentos são agrupados de acordo com suas semelhanças e co-relacionamentos. Diferentemente do que ocorre na classificação, na *Cluster*ização não existe um conhecimento anterior das classes possíveis ou existentes, A descoberta dos grupos ocorre na execução do algoritmo, baseado unicamente no conteúdo dos documentos.

A *Clusterização* de textos tem a finalidade de ordenar textos em grupos (*Clusters*), de forma que os objetos pertencentes ao mesmo *Cluster* tenha alta similaridade, ou seja, possuam características em comum, sendo diferentes dos objetos agrupados em outros *Clusters*, podendo haver formas diferentes, subjetivas, de grupos para um mesmo conjunto de dados, como mostra a Figura 1. O primeiro aglomerado, Figura 1 (b) separa em dois grupos levando em consideração a forma dos objetos. Na segunda Figura 1(c), o aglomerado separa os objetos pelo seu preenchimento. No final, o último aglomerado, Figura 1(d) divide os objetos em quatro grupos considerando as características forma e preenchimento (FACELI *et al.*, 2011).

Figura 1- Aglomerados



Fonte: FACELI *et al.*, 2011.

A técnica de KDT utilizada neste trabalho é a *Clusterização*. Esse método consiste na identificação de grupos de textos onde os mesmos têm características semelhantes aos do mesmo grupo e onde os grupos tenham características diferentes entre si (WIVES, 1999).

3.3 Métricas para Análise de Agrupamento de Texto

As métricas são utilizadas para avaliar o desempenho da mineração de textos para cada agrupamento considerado. De acordo com Halkidi *et al*. (2001), a avaliação dos agrupamentos pode ser dividida em três grandes classes de métricas: internas ou não supervisionadas; externas ou supervisionadas e relativas.

Nas métricas internas ou não supervisionadas, utilizam-se apenas informações contidas nos grupos gerados para realizar a avaliação dos resultados, ou seja, não se utilizam informações externas. As medidas mais usadas, de acordo com Guelpeli (2012), para este fim, são *Coesão, Acoplamento* e *Coeficiente de Silhouette*.

A *Coesão* (Equação 1) mede a similaridade entre os elementos do mesmo agrupamento. Quanto maior a similaridade entre eles, maior a coesão deste agrupamento (GUELPELI, 2012). *Coesão* (*C*): Equação 1:

$$C = \frac{\sum_{i>j} Sim(Pi, Pj)}{\frac{n(n-1)}{2}} \tag{1}$$

Onde *Sim* (*Pi, Pj*) é o calculo da similaridade entre os textos *i* e *j* pertencentes ao agrupamento *P, n* é o número de textos no agrupamento *P*, e *Pi* e Pj são membros do agrupamento *P* (GUELPELI, 2012).

O *Acoplamento* (Equação 2) mede a similaridade média de todos os pares de elementos, sendo que um elemento pertence a um agrupamento e o outro não pertence a esse mesmo agrupamento (GUELPELI, 2012).

Acoplamento (A): Equação 2:

$$A = \frac{\sum_{i>j} Sim(Ci, Cj)}{\frac{n_a\,(n_a - 1)}{2}} \tag{2}$$

Onde C é o centroide de determinado agrupamento, presente em *P*, *Sim (Ci, Cj)* é o cálculo da similaridade do texto i pertencente ao agrupamento *P* e o texto *j* não pertence a *P*, *Ci* centroide do agrupamento *P e Cj* é centroide do agrupamento Pi e na é o número de agrupamentos presentes em P (GUELPELI, 2012).

O *Coeficiente Silhouette* (Equação 3) baseia-se na ideia de quanto um objeto é similar aos demais membros do seu grupo, e de quanto este mesmo objeto é distante dos de outro grupo. Assim, essa medida combina as medidas de *Coesão* e *Acoplamento* (GUELPELI, 2012). *Coeficiente de Silhouette* (*S*): Equação 3:

$$S = \frac{(b(i) - a(i))}{\max\,(a(i), b(i))} \tag{3}$$

Onde *a* (*i*) é a distância média entre o *i-ésimo* elemento do grupo e os outros do mesmo grupo. O *b*(*i*) é o valor mínimo de distância entre o *i-ésimo* elemento do grupo e qualquer outro grupo, que não contém o elemento, e *max* é a maior distância entre *a* (*i*) e *b*(*i*) (GUELPELI, 2012).

Para as métricas externas ou supervisionadas, os resultados dos agrupamentos são avaliados por uma estrutura de classes pré-definidas que refletem a opinião de um especialista humano. Para esse tipo na opinião de Guelpeli (2012), são usadas medidas como: *recall*, *precision* e como medida harmônica destas duas, o *F-measure* descritas a seguir.

O *Recall* mede a proporção de objetos corretamente alocados a um agrupamento, em relação total de objetos da classe associada a este agrupamento GUELPELI (2012, apud, Manning *et al* 2008).

A *Precision* mede a proporção de objetos corretamente alocados a um agrupamento, em relação ao total de objetos deste agrupamento GUELPELI (2012, apud, Manning *et al* 2008).

O *F-Measure* é a medida harmônica entre o *Precision* e o *Recall* que, no *F-Measure*, assume valores que estão no intervalo de [0,1]. O valor zero indica que nenhum objeto foi agrupado corretamente, o valor um, que todos os objetos estão contidos correta-

mente agrupados. Assim, um agrupamento ideal deve retornar um valor igual a um GUELPELI (2012, apud, Manning *et al* 2008).

Para Guelpeli (2012), a métrica relativa tem como objetivo encontrar o melhor conjunto de grupos que um algoritmo de agrupamentos pode definir, a partir de certas suposições e parâmetros. A avaliação de um agrupamento é realizada por comparações entre esse agrupamento, gerados pelo mesmo algoritmo, mas com diferentes parâmetros de entrada.

Neste trabalho adotaram-se apenas as métricas internas ou não supervisionadas para avaliar os *clusters*, devido esse tipo de métrica utilizar apenas informações contidas nos grupos para realizar a avaliação dos resultados. O modelo Cassiopeia só trabalha com as métricas internas e devido o foco deste trabalho ser a utilização do modelo Cassiopeia como *Cluster*izador de textos, as métricas para a análise de agrupamentos textuais utilizadas serão as internas.

3.4 Contextualização

Os *tweets* coletados para formação do *corpus*[2] neste trabalho foram oriundos do Exame Nacional do Ensino Médio (ENEM). Criado em 1998, O ENEM é uma prova elaborada pelo Ministério da Educação com o objetivo de avaliar o desempenho do estudante ao fim da escolaridade básica e aferir o desenvolvimento das competências e habilidades necessárias ao exercício pleno da cidadania.

No Brasil, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), órgão vinculado ao Ministério da Educação (MEC) é responsável pela elaboração e aplicação do ENEM.

Na sua primeira edição em 1998, o ENEM contou com um número relativamente pequeno de participantes: cerca de 115.600. Não obstante, em 2008 o ENEM atingiu a marca de 4.018.050 de inscritos, alcançado patamar superior aos 4.600.000 inscritos na edição de 2010. O ENEM de 2016 registrou 9.276.328 inscritos. O resultado é o segundo maior da série histórica do exame – ficando atrás apenas da edição de 2014, quando foram registrados 9.490.952 participantes. Na edição de 2015, o número chegou a 8.478.096 inscritos[3].

Podem participar do exame alunos que estão concluindo ou que já concluíram o ensino médio em anos anteriores. O ENEM é composto por quatro provas de múltipla escolha, com 45 questões cada, e uma redação.

O ENEM tem uma série de funções[4]:

- O exame é usado como um vestibular nacional de uma série de universidades públicas. Com a nota do ENEM, o estudante pode se inscrever no Sistema de Seleção Unificada (Sisu), sistema criado pelo governo para selecionar alunos para as instituições públicas de ensino superior.
- Os estudantes também utilizam o ENEM para conseguir uma bolsa de estudos em uma universidade particular por meio do Programa Universidade Para Todos (Prouni).

---

[2] *corpus* é o conjunto de documentos sobre determinado tema vide seção 2.7.

[3] http://www.brasil.gov.br

[4] http://portal.inep.gov.br/web/enem/sobre-o-enem

Este programa do governo federal oferece bolsas de estudo parciais, de 50%, e integrais, a estudantes de baixa renda.

- Outro programa de acesso ao ensino superior que exige o ENEM é o Fundo de Financiamento Estudantil (Fies), que concede bolsas restituíveis a estudantes que não têm condições de pagar as mensalidades da graduação. O Fies funciona como um empréstimo: aluno completa o curso com bolsa, e depois de formado paga a dívida ao governo, com juros mais baixos.
- O ENEM também é necessário para os estudantes de graduação que queiram fazer um intercâmbio no exterior pelo programa Ciência sem Fronteiras.
- Com o ENEM, o candidato também consegue emitir o Certificado de Conclusão do Ensino Médio.

O exame tem abrangência nacional e, até o presente, é realizado anualmente, com conteúdos que se desdobram em quatro áreas de conhecimento, a saber: 1. Ciências Humanas e suas Tecnologias, 2. Ciências da natureza e suas Tecnologias, 3. Linguagens, códigos e suas Tecnologias e Redação, 4. Matemática e suas Tecnologias.

A área de conhecimento 1 abrange quatro componentes curriculares: *(i) História, (ii) Geografia, (iii) Filosofia e (iv) Sociologia*. A área dois envolve Química, Física e Biologia, já área três envolve Língua portuguesa, Literatura, Língua Estrangeira, Artes, Educação Física e Tecnologias da Informação e Comunicação. A quarta e última envolve apenas Matemática.

Em outubro de 2016 aconteceram movimentos que ocuparam escolas, universidades, institutos federais e outros locais em diversos estados do país. O movimento de ocupação nas escolas consistiu em uma ação desenvolvida pelos estudantes para ocupar fisicamente o espaço escolar para lutar pela melhor condução de ações do Estado.

O movimento de ocupação de escolas que teve início em outubro de 2016 perdurou até meados de dezembro do mesmo ano, e acabou interferindo no processo de aplicação das provas do ENEM 2016. Por conta das ocupações, 401 locais de aplicação das provas tiveram de ser substituídos pelo governo federal. O MEC decidiu adiar o ENEM que seria nos dias 5 e 6 de novembro para os dias 3 e 4 de dezembro para os estudantes inscritos que fari-

am as provas em escolas que estavam ocupadas. O adiamento atingiu 191.494 candidatos de 304 locais em 20 estados[5].

A aplicação das provas do ENEM ocorreu em dois momentos. Nos dias 5 e 6 de novembro para estudantes que fizeram provas em escolas que não tinha manifestações e nos dias 3 e 4 de dezembro para estudantes que fariam as provas em escolas que estavam ocupadas. Assim, foram realizadas coletas de *tweets* sobre a realização da prova do ENEM nos dois períodos de aplicação das provas.

No ano de 2016 também houve o lançamento de uma ferramenta inédita e gratuita de preparação para as provas do ENEM: a Plataforma Hora do ENEM. O Hora do ENEM foi um projeto idealizado pelo MEC, voltado para estudantes que iriam fazer o Exame[6].

A iniciativa reúne na *internet* um conjunto de ações, como simulados e vídeo aulas, para auxiliar na preparação dos estudantes. Ao usar um computador, tablet ou celular para acessar a plataforma, o estudante marca o curso em que quer passar e quanto tempo tem para estudar por dia para o ENEM. A partir desse diagnóstico, a plataforma oferece um plano de estudo com pontos fortes e fracos na medida do participante, com exercícios, resumos e videoaulas direcionados. A plataforma ainda traz ao estudante boletim de notícias diário com informações sobre o ENEM, programa televisivo com dicas para as áreas de conhecimento, vídeos com resoluções de questões que caíram em anos anteriores da avaliação, entre outras novidades.

Os estudantes se inscrevem gratuitamente na plataforma de estudos. Ao longo do ano de 2016, foram aplicados quatro simulados-testes com a mesma metodologia de elaboração das questões do ENEM. Foram realizadas coletas de *tweets* sobre a realização de dois dos quatro simulados-testes aplicados pelo Hora do ENEM.

Os usuários do *twitter* tendem a expressar livremente as opiniões comuns sobre todo o processo relacionado ao ENEM, assim há uma ampla e valiosa gama de informações que pode ser explorada para a descoberta de conhecimento através de técnicas de mineração de textos. No presente trabalho foram utilizados os textos provenientes do ENEM 2016, visando a descoberta de conhecimento inserida dentro do contexto educacional. O ENEM é

[5] http://portal.mec.gov.br

[6] http://horadoenem.mec.gov.br

considerado não apenas um indicador de qualidade para o ensino médio, mas também um dos instrumentos de política pública voltado a permitir maior democratização das oportunidades de acesso ao ensino superior.

3.5 *Twitter*

O *Twitter* é uma rede social que permite a seus usuários transmitir uma informação através de mensagens de textos rápidas. As mensagens são limitadas a 140 caracteres conhecidas como *tweets*.

Sua estrutura dinâmica permite que qualquer usuário tenha acesso às informações que são constantemente postadas, sem que para isso restrinja aos usuários possuírem alguma permissão de conexão entre eles. Russel (2013) aponta que esse é o grande diferencial do *Twitter* em relação a outras redes sociais populares, tornando o *Twitter* uma rede social mais interessante de se explorar.

Na rede social *Twitter*, os *tweets* podem ser agrupados por *hashtags* que são palavras precedidas pelo caractere #, utilizado para marcar palavras-chave ou tópicos em um *tweet*. As *hashtags* são utilizadas para categorizar os conteúdos publicados nas redes sociais, ou seja, criar uma interação dinâmica do conteúdo com os outros integrantes da rede social, que estão ou são interessados no respectivo assunto publicado. Além disso, os usuários podem dar *retweet* em um *tweet* publicado por outros usuários. Um *retweet* é uma nova postagem no *tweet* de alguém[7].

O *Twitter* foi utilizado como fonte de exploração de dados para este trabalho, por ser um meio eletrônico de colaboração, comunicação e troca de ideias entre usuários que possuem interesses em comum. Com uma grande quantidade de usuários ativos, possui alcance global, que induz seus usuários a compartilhar suas ideias constantemente, gerando grande quantidade de informação a cada instante.

[7] https://support.*twitter*.com

Além disso, o *Twitter* fornece uma API[8] que permite a recuperação de postagens de usuários. Para realizar a extração dos dados textuais do *Twitter*[9], o usuário pode optar por duas opções de pesquisa disponíveis no sistema, a busca por *tweets* recentes e a busca em tempo real. Através da busca por *tweets* recentes, é possível recuperar mensagens postadas há até 7 dias anteriores a data da busca. Já a busca por *tweets* em tempo real permite que os textos sejam capturados à medida que são lançados na base de dados do *Twitter* e disponibilizados na *web*.

Sendo assim, este trabalho faz uso de ferramentas para coletar campos textuais ou *tweets* do servidor do *Twitter* para fins de aprendizagem eletrônica e, além disso, identificar informações explícitas nas discussões dos usuários.

Os textos utilizados neste trabalho (ENEM 2016), se encontram disponíveis na *web*, sob licença aberta, estruturados de maneira legível para máquinas e utilizando um formato não proprietário (*csv*).

3.6 Ferramentas Utilizadas

Nas seções seguintes são abordadas as ferramentas utilizadas para a coleta de campos textuais do *Twitter*, ou seja, utilizadas para a recuperação da informação. São *softwares* que percorrem sítios da internet como intuito de coletar automaticamente os dados destes. Após a recuperação destes dados pretendidos para a análise, é possível criar um *corpus* que servirá de base para aplicar as técnicas de mineração de textos. Portanto, a etapa de recuperação e coleta de dados tem como função a criação de uma base de dados textual chamada *corpus* ou corpora. Um *corpus* nada mais é que uma coleção de textos (MANNING et al. 2008).

---

[8] A sigla API refere-se ao termo em inglês "*Application Programming Interface*" significa "Interface de Programação de Aplicativos" -

[9] Twitter Developer Documentation. Disponível em**:** <https://dev.*twitter*.com/rest/public>

3.6.1 *Your Twapper Keeper* (Ytk)

O *Your Twapper Keeper*[10] é uma ferramenta que permite aos usuários arquivar, organizar e analisar os *tweets* com base em #hashtags, perfis ou palavras específicas. Muito usado por organizações acadêmicas para acompanhar, monitorar, e analisar o tráfego do *Twitter*. Esta ferramenta possibilita arquivar dados do *Twitter* em tempo real diretamente no servidor de destino, para compor a base de documentos a ser trabalhada na etapa seguinte. O *Your Twapper Keeper* é um *software* de código aberto e gratuito, seu uso é simples. Basta fazer o login com uma conta do *twitter* e inserir o termo que deseja pesquisar no campo *Keyword or Hashtag.*

Desenvolvido pela equipe que trabalhava no TwapperKeeper.com, está disponibilizado por John O'Brien III. Porém, devido a sua característica de código aberto, o Media Lab da Universidade Federal do Rio de Janeiro - UFRJ desenvolveu uma adaptação de exportação em arquivos csv.

3.6.2 *Collect Convert*

O *Collect Convert*[11] é um *Script* (conjunto de instruções para que uma função seja executada em determinado aplicativo) para coleta automatizada de texto e imagem desenvolvido no Laboratório de estudos sobre Imagem e Cibercultura (Labic) da Universidade Federal do Espírito Santo –UFES. Funciona a partir do acesso ao código fonte de uma página *web* que está na linguagem HTML (HyperText Markup Language, que significa Linguagem de Marcação de Hipertexto, é uma linguagem de marcação utilizada na construção de páginas *Web*) e consegue coletar *tweets* de até 7 dias anteriores. A ferramenta está disponível para download.

---

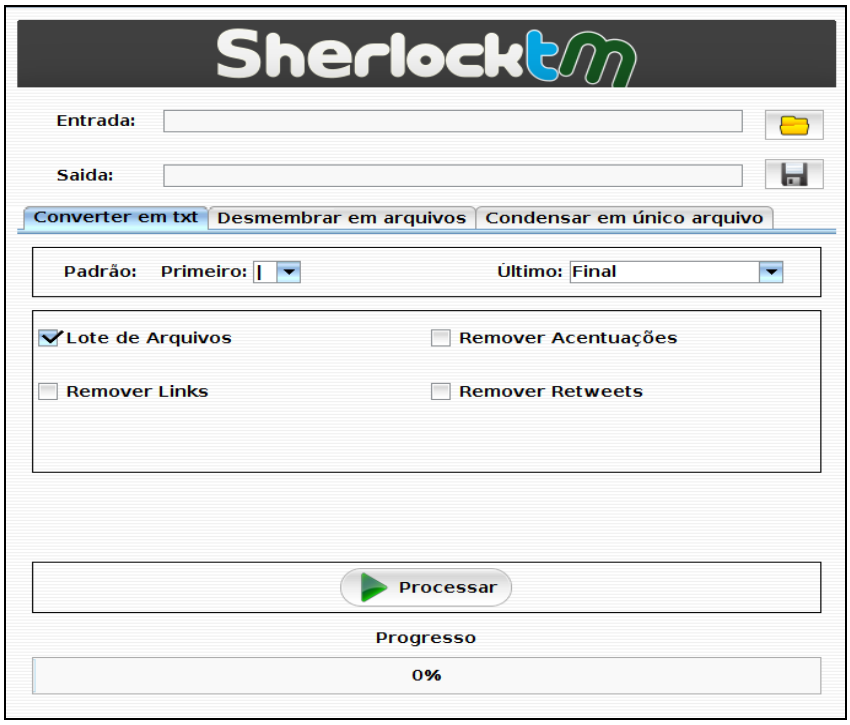
[10] YourTwapperKeeper. Disponível em: <https://github.com/540co/yourTwapperKeeper>. Acesso em: 12 de Janeiro 2017.

[11] *Collect Convert*, ferramenta disponibilizada na página do laboratório Labic Ufes. Disponível em: <https://github.com/ufeslabic >. Acesso em: 20 de Janeiro 2017.

3.6.3 *Sherlocktm*

É uma ferramenta desenvolvida em linguagem Java (linguagem de programação interpretada orientada a objetos), destinado à área de mineração de textos. É uma ferramenta mais completa, pois integra um grupo de funções como coleta automatizada de *tweets*, limpeza, organização e conversão dos arquivos para diferentes formatos. O *SherlockTM* foi desenvolvido por um integrante do laboratório do grupo de pesquisa Mineração de Texto, Processamento de Linguagem Natural e Aprendizado de Máquina - MTPLNAM da Universidade Federal dos vales do Jequitinhonha e Mucuri - UFVJM. A Figura 2 mostra a interface do *SherlockTM*.

Figura 2 - Interface do SherlockTM



Fonte: Própria autora

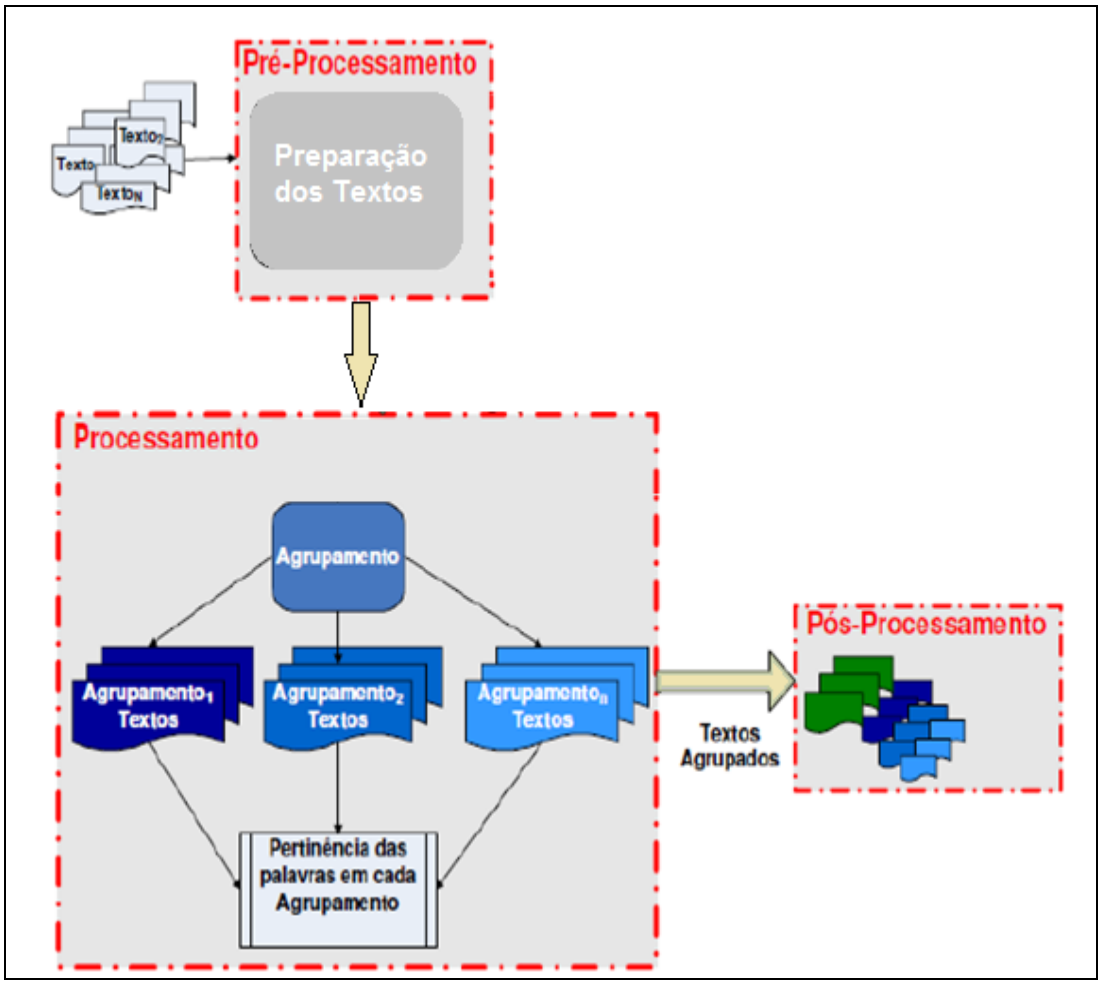
3.6.4. Modelo Cassiopeia

Através de estudos (Guelpeli 2012), desenvolveu um modelo de agrupamento denominado Cassiopeia. O Cassiopeia é um *Cluster*izador de texto, que utiliza um algoritmo de mineração de textos, que tem como principal finalidade gerar agrupamentos, ou seja, *clusters* (grupos) de documentos textuais que apresentam algum tipo similaridade. O Cassiopeia é independente do idioma e do domínio, trazendo um diferencial e um ganho com relação a outros métodos de agrupamento de texto encontrados na literatura.

O modelo Cassiopeia foi criado para possibilitar o agrupamento dos textos, com maior qualidade nas avaliações em domínios distintos, independentes do idioma, avaliados pelas métricas internas. Este modelo utiliza no processamento a seleção de atributos e o processo de agrupamento hierárquico de texto e no pós-processamento são apresentados os agrupamentos por similaridade e os textos sumarizados (GUELPELI, 2012).

Neste trabalho o modelo Cassiopeia é utilizado para realizar o processo de mineração dos *tweets* e criar os grupos similares de *tweets* ou *Clusters*. A Figura 3 ilustra a metodologia do Cassiopeia.

Figura 3 - Metodologia do Modelo Cassiopeia



Fonte: GUELPELI, (2012) (Adaptado)

O Modelo Cassiopeia possui o seu processo dividido em três etapas distintas. A etapa de pré-processamento, processamento e a de pós-processamento.

O Pré-processamento é a etapa realizada após a Coleta, com o objetivo de se obter alguma estrutura para a formação do *corpus*. O principal objetivo de pré-processar um texto, consiste na filtragem e limpeza dos dados, eliminando redundâncias e informações desnecessárias para o conhecimento que se deseja extrair (GONÇALVES et al., 2006).

Depois dos *tweets* limpos, pré-processados e transformados de acordo com a proposta, começa a fase de processamento, que usa o processo de *Cluster*ização, cuja finalidade é juntar os *tweets* automaticamente por similaridade. Os agrupamentos são criados com a utilização de um *Cluster*izador automático, de forma que independente do seu domínio e idio-

ma possa ser realizado o agrupamento, trazendo um diferencial e um ganho com relação a outros métodos de agrupamento de texto encontrados na literatura.

Assim, na fase de processamento, o *Cassiopeia* utiliza o processo *Cluster*izador para analisar o *corpus*, encontrar similaridades entre os *tweets* e gerar *Clusters* que serão analisados para gerar conhecimento. O *Cassiopeia* recebe um arquivo no formato.txt com as amostras de *tweets* dos períodos definidos. Estas amostras serão processadas e os *tweets* serão separados automaticamente por similaridade.

Com o resultado disponível, já se pode analisar, interpretar ou avaliar o conhecimento descoberto. Parte-se para a etapa de pós-processamento, na qual cada agrupamento ou *Cluster* terá, por similaridade, um conjunto de *tweets*, que têm alto grau de informatividade. Cada *Cluster* será analisado gerando informações específicas sobre o ENEM.

Na etapa de pós-processamento, na qual cada um dos agrupamentos ou subagrupamentos terá, por similaridade, um conjunto de textos, com alto grau de informatividade, o Cassiopeia fará ainda a mensuração dos resultados com a aplicação das métricas internas (coesão, acoplamento e coeficiente *silhouette*).

Finalizado este processo, parte-se para a etapa de avaliação do resultado da Mineração de Textos, onde cada *Cluster* será analisado, gerando informações específicas sobre os mesmos.

3.7 *Corpus*

A palavra *corpus* é de origem latina, e significa corpo; no contexto acadêmico, *corpus* é o conjunto de documentos sobre determinado tema. Segundo Bauer e Aarts (2002), o *corpus* de um tema é composto pelos materiais identificados como fontes importantes para que o aluno/pesquisador possa fundamentar seu texto, adequado ao caráter científico necessário à sua pesquisa.

Segundo Sinclair (2005), o *corpus* ou corpora é uma coletânea de textos em certo idioma que está já em formato eletrônico. Especificamente, esses textos devem ser selecionados de acordos com critérios externos, ou seja, critérios que nascem a partir das ne-

cessidades da pesquisa na qual o *corpus* será usado e que sejam capazes de representar uma língua ou uma parcela de língua.

A necessidade de criação do primeiro *corpus* se deu no ano de 1964 quando o *Brown University Standard Corpus of Present-Day American English* continha uma grande quantidade de dados para informatizar. A criação do primeiro *corpus* linguístico eletrônico, o *corpus Brown,* impulsionou o desenvolvimento da área conhecida atualmente por Linguística de *Corpus*, uma das áreas de pesquisa de linguagem mais ativa nos últimos anos (SARDINHA, 2000).

Uma coletânea de textos, quando caracteriza como *corpus*, é inegavelmente uma fonte de conhecimento não estruturado que permite a extração de dados linguísticos reais e em larga escala (ALUÍSIO; ALMEIDA, 2006).

De acordo com Trask (2004), "a partir de *corpora*, podem-se fazer observações precisas sobre o real comportamento linguístico de falantes reais, proporcionando informações altamente confiáveis e isentas de opiniões e de julgamentos prévios sobre os fatos de uma língua".

Desta forma, por meio do *corpus,* podem-se observar aspectos bastante relevantes para uma pesquisa.

4 METODOLOGIA

Conforme mencionado anteriormente, o objetivo da pesquisa é analisar a opinião dos usuários do *Twitter* sobre o Exame Nacional do Ensino Médio (ENEM) 2016 através da técnica de Mineração de Textos usando *Clusterização* para extrair conhecimento inerente aos textos analisados. Para cumprimento do propósito da pesquisa, foi utilizada a metodologia mista, quantitativa e qualitativa. Creswell e Plano Clark (2011 apud PARANHOS, RANULFO et al. 2016) definem métodos mistos como um procedimento de coleta, análise e combinação de técnicas quantitativas e qualitativas em um mesmo desenho de pesquisa.

Para construção da fundamentação teórica, foi realizada a pesquisa bibliográfica, visando o desenvolvimento dos assuntos foco do trabalho, como Mineração de Textos e *Clusterização.*

Dessa maneira, a pesquisa ocorreu envolventos etapas de caráter quantitativo e qualitativo. A etapa quantitativa teve como objetivo coletar *tweets* da rede social *Twitte*r para recuperação dos dados textuais para formação do *corpus* no domínio educacional ENEM 2016, que permitiu construir o corpus a ser *clusterizado* através do Modelo Cassiopeia. A etapa da pesquisa, de natureza qualitativa, buscou compreender, por meio de análises o resultado do processamento, ou seja, buscou conhecer quais as informações presentes nos agrupamentos.

A metodologia deste trabalho se divide em:

- Coleta dos *tweets*;
- Pré-processamento da base de dados textuais;
- Transformação dos dados textuais;
- Clusterização dos dados como Modelo Cassiopeia;
- Análise dos resultados.

As subseções a seguir mostram de forma detalhada cada um dos passos necessários para a execução desse trabalho.

4.1 Coleta dos *Tweets*

Na etapa de extração ocorre a coleta onde se utiliza *web crawlers,* programas que visitam sítios e extraem informações, para buscar os textos que serão utilizados para a extração de conhecimento. Segundo Aranha (2007) coletar dados é uma atividade trabalhosa, um dos motivos é que os dados podem não estar disponíveis em um formato apropriado para serem utilizados no processo de mineração de textos.

A coleta dos *tweets* para a mineração de texto ocorreu através de três ferramentas de coleta: *Your Twapper Keeper, Collect Convert e SherlockTM*, ou seja, capturando os comentários à medida em que eles eram postados no momento da coleta. O *corpus* produzido neste trabalho constituiu-se de *tweets* sobre o tema ENEM no ano de 2016 (SILVA; GUELPELI, 2017). Os *tweets* foram coletados de acordo com o cronograma do ENEM 2016/2017.

As coletas foram realizadas a partir do período de inscrições do ENEM 2016 (09 de Maio de 2016) e encerradas no dia de divulgação dos resultados (18 de janeiro de 2017), para que assim, fosse feita uma análise dos acontecimentos relacionados ao evento durante todo o seu período de realização, totalizando 239.622 *tweets.* O *corpus* foi dividido em categorias de acordo com o período das coletas que foram: Maio ENEM 2016, Agosto ENEM 2016, Setembro ENEM 2016, Outubro ENEM 2016, Novembro ENEM 2016, Dezembro ENEM 2016 e Janeiro ENEM 2016.

As coletas realizadas sobre o ENEM 2016 foram agrupadas em sete categorias, como mostra a Tabela 1:

Tabela 1 – Coletas/Categorias do ENEM 2016

Inscrições do ENEM	9 a 20 de Maio
3º Simulado Online	03 a 11 de setembro
4º simulado Online	08 a 23 de Outubro
1ª Aplicação das Provas	5 e 6 de Novembro
2ª Aplicação das Provas	3 e 4 de Dezembro
Aplicação Prova Pessoas Privadas de Liberdade	13 e 14 de Dezembro
Resultado do ENEM	18 de janeiro de 2017

Fonte: Própria autora

A Tabela 2 mostra os meses e os termos das coletas que foram realizadas. Dentro de cada mês, foram realizadas coletas sobre o termo geral "ENEM 2016", "#enem 2016", datas específicas de eventos do ENEM, ou de acordo com os acontecimentos relacionados ao ENEM. Foram realizadas coletas do termo "#enem 2016", pois hashtags são muito utilizadas em redes sociais para categorizar os conteúdos publicados nas redes sociais.

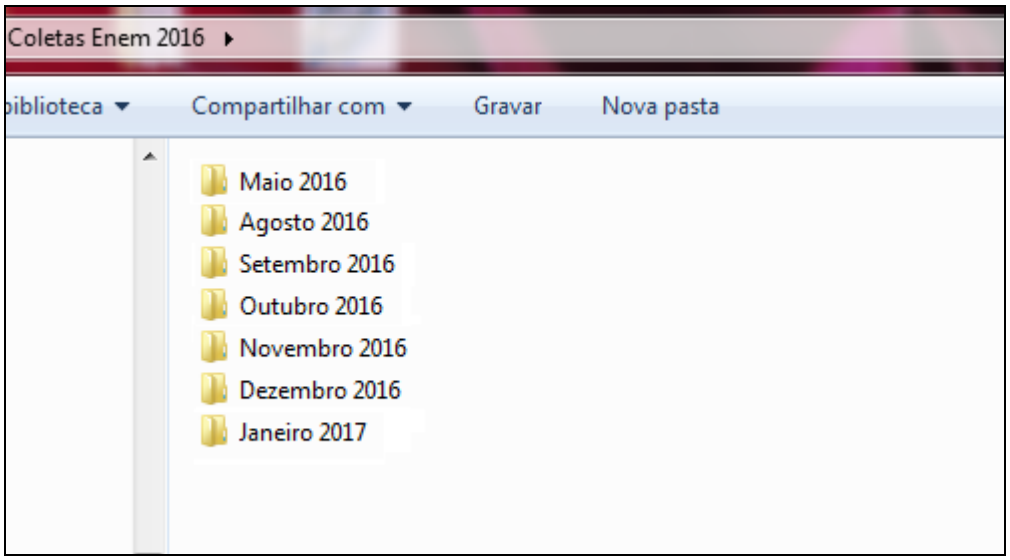
Tabela 2 - Termos de Coleta sobre o ENEM 2016

Maio	Termo geral "ENEM 2016" e "Inscrição ENEM 2016" (9 a 20 de Maio)
Agosto	Termo geral "ENEM 2016" e "#enem 2016"
Setembro	Termo geral "ENEM 2016", "#enem 2016" e "3º Simulado Online" (03 a 11 de setembro)
Outubro	Termo geral "ENEM 2016", "#enem 2016" e "4º Simulado Online" (08 a 23 de Outubro)
Novembro	Termo geral "ENEM 2016", "#enem 2016", "Adiamanto ENEM 2016", "Cancelamento ENEM 2016", "Fraude ENEM 2016", "Problemas ENEM 2016" "Redação ENEM 2016", "Segurança ENEM 2016", "Ocupação Escolas" e 1ª Aplicação das Provas (5 e 6 de Novembro)
Dezembro	Termo geral "ENEM 2016", "#enem 2016", "Cartão Confirmação ENEM 2016", "Problemas ENEM 2016","Cancelamento ENEM 2016", " Adiamento ENEM 2016 ", "2ª Aplicação das Provas" (3 e 4 de Dezembro), "Fraude ENEM 2016", "Redação ENEM 2016", "Vazamento ENEM 2016" e "Aplicação Prova para Pessoas Privadas de Liberdade" (13 e 14 de Dezembro)
Janeiro	Termo geral "ENEM 2016", "#enem 2016" e "Resultado do ENEM" (18 de janeiro de 2017)

Fonte: Prórpia autora

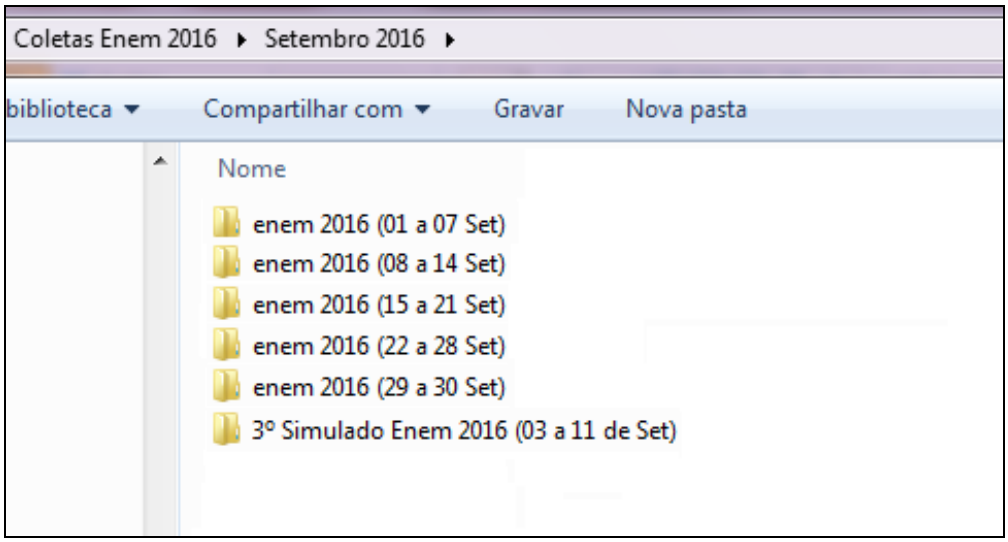
As coletas/categorias foram separadas em pastas compondo sete categorias de acordo com o mês correspondente. O *corpus* foi dividido em subpastas, conforme exemplifica as Figuras 4 e 5.

Figura 4 - Pastas das Coletas dos *Tweet*



Fonte: Própria autora

Figura 5 - Subpastas com os Termos e as Datas das Coletas



Fonte: Própria autora

Nos meses de abril e junho de 2016, quando aconteceram o 1º e o 2º simulados *online*, não foi possível realizar a coleta de *tweets*. Isso ocorreu pelo fato da plataforma que realizou os simulados, a Hora do ENEM, ser uma ferramenta inédita, e ainda não conhecida. Neste período ainda estavam sendo levantadas informações a respeito da mesma e só mais

adiante dos estudos identificamos a necessidade de realizar coletas nos períodos de aplicação dos simulados. Assim, dos quatro simulados *online* realizados, as coletas realizadas foram referente ao 3º e 4º simulados.

A princípio a ferramenta utilizada pararealizar as coletas foi a *Your Twapper Keeper* que possibilita arquivar dados do *Twitter* em tempo real diretamente no servidor de destino, para compor a base de documentos a ser trabalhada na etapa seguinte. Porém, o funcionamento desta ferramenta é em tempo real e quando acontece alguma queda de energia o processo é interrompido e os *tweets* são perdidos.

Assim, passa-se a realizar as coletas com o *collect convert* que é um *Script* de coleta automatizada de texto e imagem desenvolvido no Labic-UFES[12] que funciona a partir do acesso ao código fonte de uma página *web* e consegue coletar *tweets* de até 7 dias anteriores.

Ao final do período das coletas foi realizada mais uma transição de ferramenta de coleta de *tweets*. As coletas passaram a ser realizadas pela *SherlockTM* (*Sherlock Text Mining*). A transição para o *SherlockTM* neste trabalho se justifica por ser uma ferramenta mais completa, pois além de realizar a coleta dos *tweets*, também possui funções auxiliares que ajudam no tratamento das informações, disponibilizando opções de pré-processamento e de organização, mantendo as informações em diretórios organizados hierarquicamente.

Com o *SherlockTM* além de reduzir o tempo de pré-processamento, houve melhoria na qualidade do resultado uma vez que o *software* elimina os problemas relativos a possíveis erros humanos, sendo um diferencial em relação as demais ferramentas. Consegue fornecer bons resultados de forma consistente e contínua aumentando a qualidade do processo.

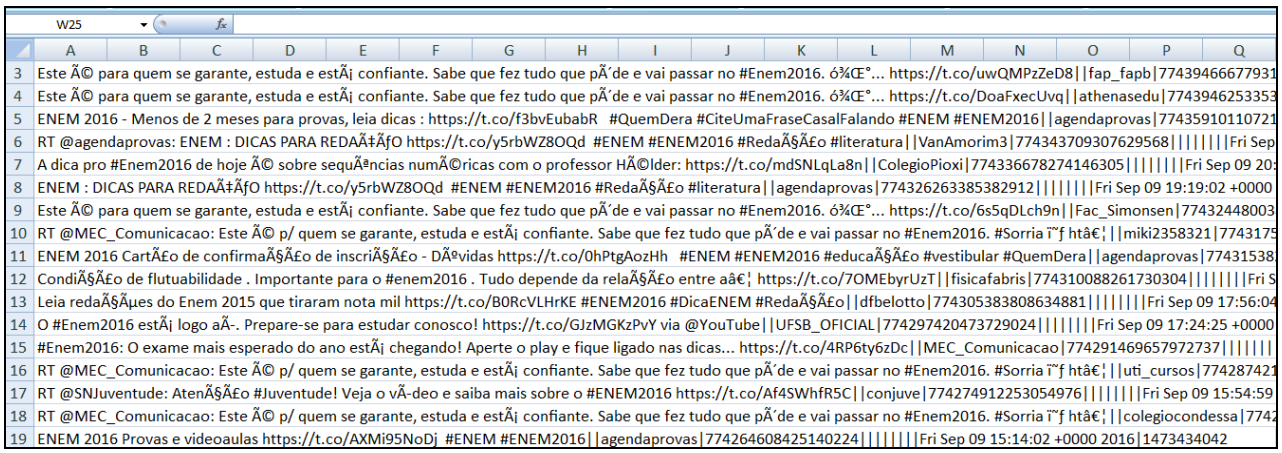
É importante destacar que as mudanças de ferramenta ocorreram na tentativa de tornar o processo mais eficaz e para obter maior qualidade no resultado final e principalmente, sem perda de *tweets*.

[12] https: //github.com/ufeslabic

4.2 Pré-Processamento da Base de Dados Textuais

Cada *tweet* foi coletado possui atributos como data e hora da criação, id, texto, idioma, nome do usuário, localidade, entre outros, como mostra a Figura 6. Porém, nem todos esses atributos serão necessários para a realização da mineração de texto. A etapa de pré-processamento será responsável por selecionar os atributos relevantes para o processamento dos textos na fase seguinte, bem como segmentá-los e remover palavras consideradas irrelevantes para a realização de uma análise de texto.

Figura 6 - *Tweet* e seus Atributos como Coletados



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
3	Este Ã© para quem se garante, estuda e estÃ¡ confiante. Sabe que fez tudo que pÃ´de e vai passar no #Enem2016. ó¾Œ°... https://t.co/uwQMPzZeD8||fap_fapb|77439466677931																
4	Este Ã© para quem se garante, estuda e estÃ¡ confiante. Sabe que fez tudo que pÃ´de e vai passar no #Enem2016. ó¾Œ°... https://t.co/DoaFxecUvq||athenasedu|77439462533533																
5	ENEM 2016 - Menos de 2 meses para provas, leia dicas : https://t.co/f3bvEubabR #QuemDera #CiteUmaFraseCasalFalando #ENEM #ENEM2016||agendaprovas|77435910110721																
6	RT @agendaprovas: ENEM : DICAS PARA REDAÃ‡ÃƒO https://t.co/y5rbWZ8OQd #ENEM #ENEM2016 #RedaÃ§Ã£o #literatura||VanAmorim3|774343709307629568|||||||Fri Sep																
7	A dica pro #Enem2016 de hoje Ã© sobre sequÃªncias numÃ©ricas com o professor HÃ©lder: https://t.co/mdSNLqLa8n||ColegioPioxi|774336678274146305|||||||Fri Sep 09 20:																
8	ENEM : DICAS PARA REDAÃ‡ÃƒO https://t.co/y5rbWZ8OQd #ENEM #ENEM2016 #RedaÃ§Ã£o #literatura||agendaprovas|774326263385382912|||||||Fri Sep 09 19:19:02 +0000																
9	Este Ã© para quem se garante, estuda e estÃ¡ confiante. Sabe que fez tudo que pÃ´de e vai passar no #Enem2016. ó¾Œ°... https://t.co/6s5qDLch9n||Fac_Simonsen|77432448003																
10	RT @MEC_Comunicacao: Este Ã© p/ quem se garante, estuda e estÃ¡ confiante. Sabe que fez tudo que pÃ´de e vai passar no #Enem2016. #Sorria ï˜ƒ htâ€¦||miki2358321|7743175																
11	ENEM 2016 CartÃ£o de confirmaÃ§Ã£o de inscriÃ§Ã£o - DÃºvidas https://t.co/0hPtgAozHh #ENEM #ENEM2016 #educaÃ§Ã£o #vestibular #QuemDera||agendaprovas|77431538																
12	CondiÃ§Ã£o de flutuabilidade . Importante para o #enem2016 . Tudo depende da relaÃ§Ã£o entre aâ€¦ https://t.co/7OMEbyrUzT||fisicafabris|774310088261730304|||||||Fri S																
13	Leia redaÃ§Ãµes do Enem 2015 que tiraram nota mil https://t.co/B0RcVLHrKE #ENEM2016 #DicaENEM #RedaÃ§Ã£o||dfbelotto|774305383808634881|||||||Fri Sep 09 17:56:04																
14	O #Enem2016 estÃ¡ logo aÃ-. Prepare-se para estudar conosco! https://t.co/GJzMGKzPvY via @YouTube||UFSB_OFICIAL|774297420473729024|||||||Fri Sep 09 17:24:25 +0000																
15	#Enem2016: O exame mais esperado do ano estÃ¡ chegando! Aperte o play e fique ligado nas dicas... https://t.co/4RP6ty6zDc||MEC_Comunicacao|774291469657972737||||||																
16	RT @MEC_Comunicacao: Este Ã© p/ quem se garante, estuda e estÃ¡ confiante. Sabe que fez tudo que pÃ´de e vai passar no #Enem2016. #Sorria ï˜ƒ htâ€¦||uti_cursos|774287421																
17	RT @SNJuventude: AtenÃ§Ã£o #Juventude! Veja o vÃ-deo e saiba mais sobre o #ENEM2016 https://t.co/Af4SWhfR5C||conjuve|774274912253054976|||||||Fri Sep 09 15:54:59																
18	RT @MEC_Comunicacao: Este Ã© p/ quem se garante, estuda e estÃ¡ confiante. Sabe que fez tudo que pÃ´de e vai passar no #Enem2016. #Sorria ï˜ƒ htâ€¦||colegiocondessa|7742																
19	ENEM 2016 Provas e videoaulas https://t.co/AXMi95NoDj #ENEM #ENEM2016||agendaprovas|774264608425140224|||||||Fri Sep 09 15:14:02 +0000 2016|1473434042																

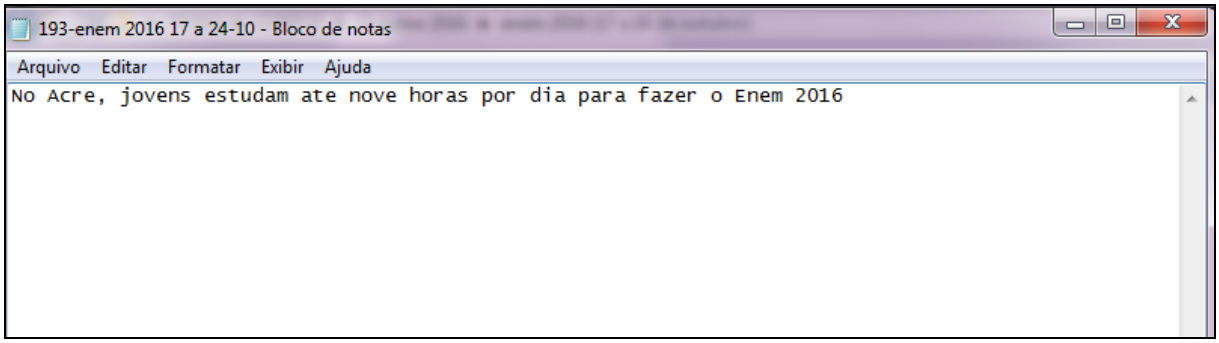
Fonte: Própria autora

O objetivo desta etapa é eliminar os textos que não se adequam às informações centrais. Consiste em transformar o conjunto de textos em um formato propício para serem submetidos ao processo de mineração (Aranha, 2007). Nesta fase é realizada a identificação de itens (características, palavras) relevantes nos documentos. Tais itens devem ser *extraídos* e organizados em dados (tabelas ou *templates*) que possam ser utilizados pelo processo de mineração.

Assim, ocorreu a limpeza e a formatação do *corpus* para o processamento computacional. Foram descartados conteúdos considerados irrelevantes para o processo de Mineração do Texto desse trabalho, como *links*, acentos, *retweets*, nomes de usuário (no *Twitter*, marcados pelo caractere '@') e geolocalização. Trabalhou-se apenas com o *tweet* em si, pois os *retweets* e demais atributos não foram relevantes para a pesquisa.

Utilizou-se nesta etapa a ferramenta *SherlockTM* tornando o pré-processamento mais rápido e consistente. No processo manual, o tempo para manipulação dos textos seria muito maior e suscetível ao erro. Com um sistema automatizado o processo ganhou agilidade na análise, limpeza e organização dos *tweets*. O resultado do pré-processamento dos *tweets* pelo *SherlockTM,* foram arquivos em formato ".txt" (que é o formato compatível para o processamento computacional do Cassiopeia), que contém a informação de cada *tweet* separadamente como mostra a Figura 7.

Figura 7 - *Tweet* após ser Pré-processado



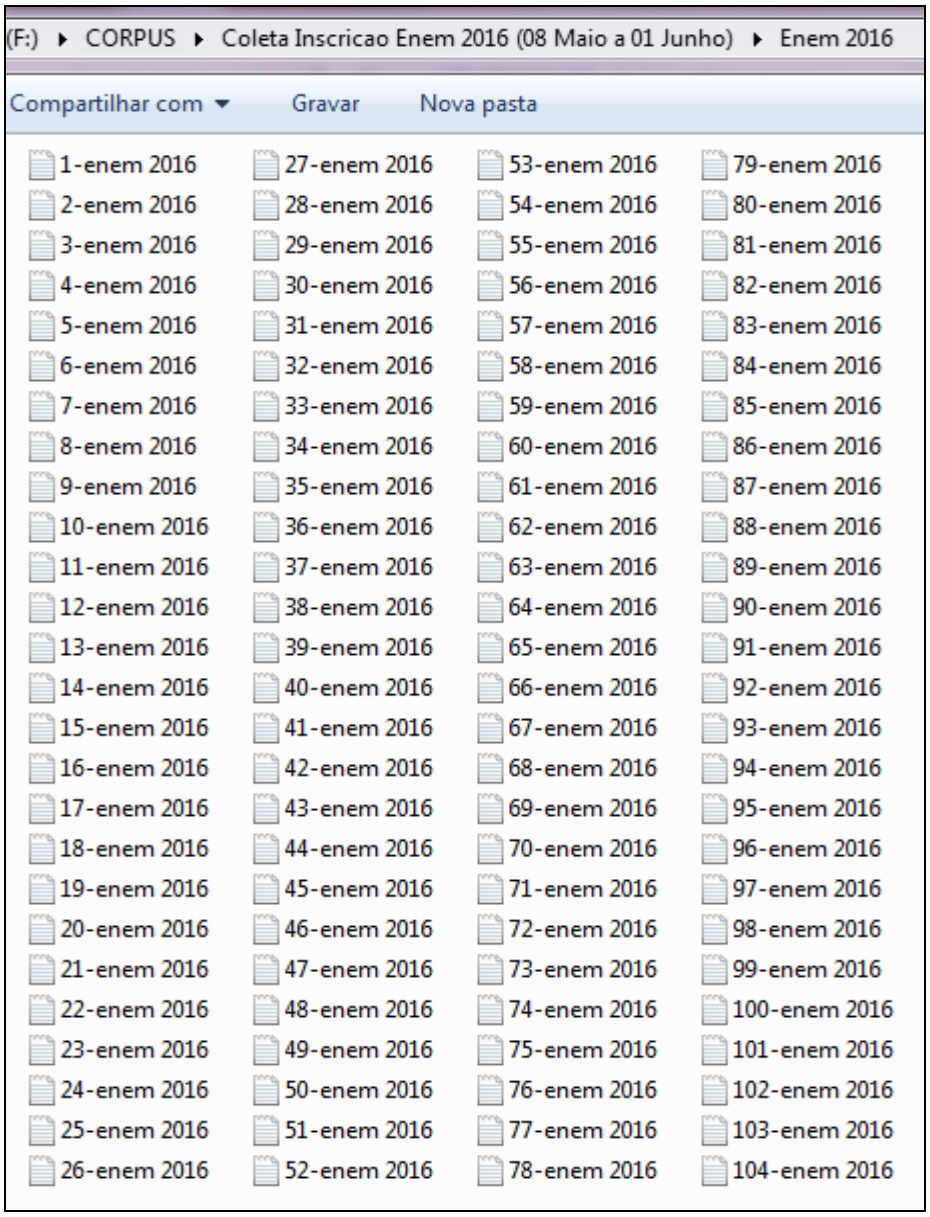
Fonte: Própria autora

4.3 Transformação dos Dados Textuais

A transformação é a etapa onde conforme explicado por Aranha (2007), os textos são catalogados, segundo um critério, para que sejam recuperados de forma mais rápida e precisa. Nesta etapa os dados deverão ser armazenados adequadamente para facilitar na utilização das técnicas de mineração de textos.

Com a finalidade de organizar e padronizar os arquivos, eles foram nomeados de acordo com as pastas em que estão inseridos, separados e colocados numa numeração sequencial. Esse padrão é mostrado na Figura 8.

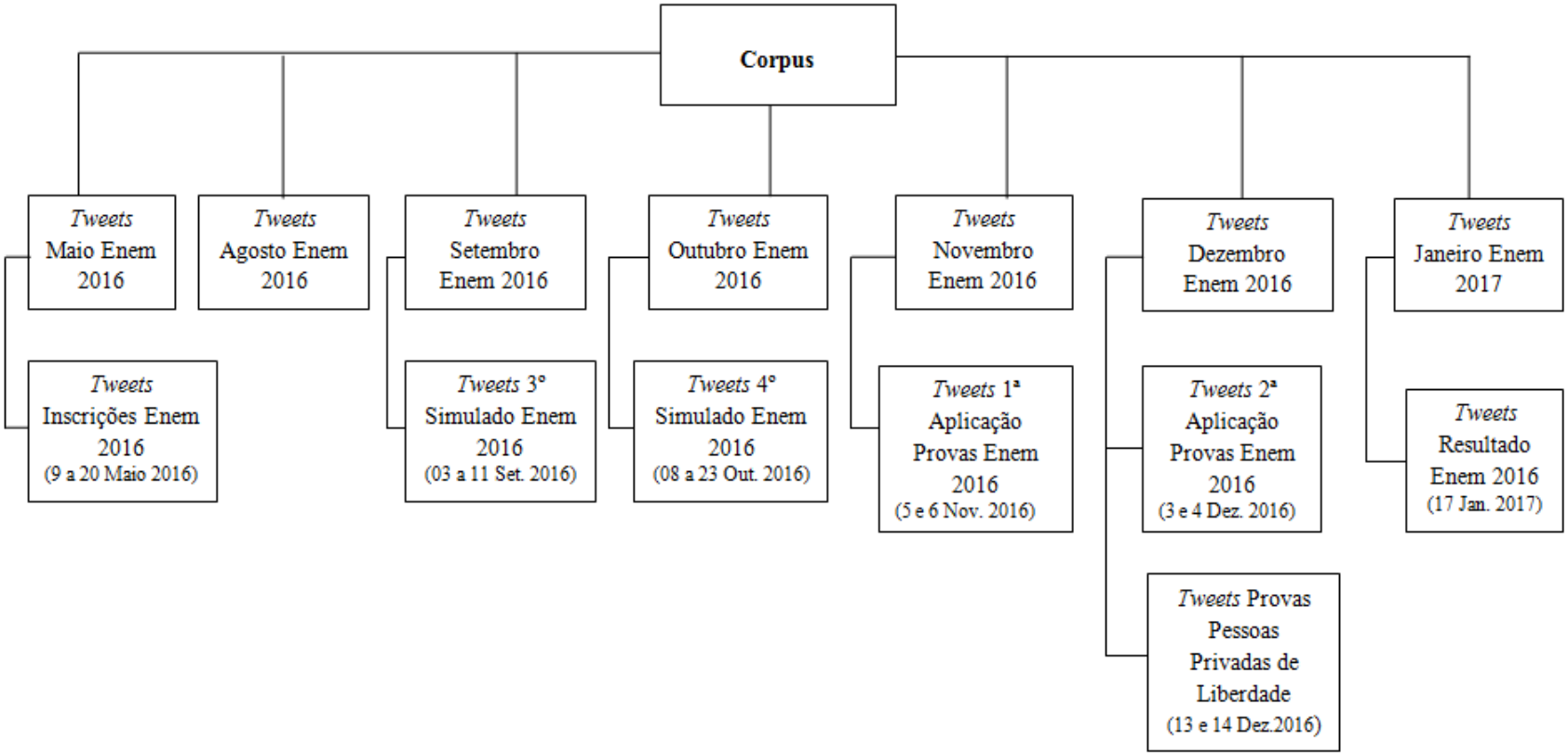
Figura 8 - Arquivos Organizados na Pasta



Fonte: Própria autora

Na figura 9, será apresentada a estrutura organizacional dos *tweets* que compõem o *corpus* formado para este trabalho. A descrição dos períodos e datas das coletas foi realizada para permitir melhor compreensão do problema em análise. Todos os textos pertencem ao domínio ENEM 2016 conforme descrito anteriormente.

Figura 9 - Diagrama do *Corpus* Utilizado Neste Trabalho



Fonte: Própria autora

Foi realizada a estatística do *corpus*, como mostra a Tabela 3 referente a dois tipos de análises, conforme trabalho de Silva e Guelpeli (2017). A primeira é referente ao *corpus* de *tweets* integral, ou seja, completa como os *tweets* foram coletados e outra para o *corpus* de *tweets* pré-processados, ou seja, após a limpeza e manipulação. As linhas têm valores relacionados à quantidade de Caracteres (símbolo, letra ou número), Caracteres e Espaços, Palavras, Palavras e Números, Porcentagem de Números, Frases e Porcentagem de Frases Repetidas, e as colunas os valores correspondentes ao *Corpus* Integral, ao *Corpus* Pré-processado e por fim a diferença entre os dois. É importante observar que houve uma redução do *corpus* integral em relação ao *corpus* pré-processado, proporcionando um ganho qualitativo e quantitativo para o processo computacional. No *corpus* Integral há um texto com 9.076.846 caracteres, já no *corpus* Pré-processado são 6.817.768 caracteres, havendo uma diferença de 2.259.078 caracteres.

Tabela 3 - Análise Estatística dos *Tweet* ENEM 2016, composta por 2 arquivos: *Corpus* de *Tweet* Integral e *Corpus* de *Tweet* Pré-processado

Arquivos	Caracteres	Caracteres e Espaços	Palavras	Palavras e Números	Porcentagem de Números	Frases	Porcentagem de Frases Repetidas
***Corpus* Integral**	9.076.846	10.650.837	1.567.559	1.891.997	17,15%	151.842	58,34%
***Corpus* Pré-processado**	6.817.768	8.183.776	1.210.742	1.497.214	19,13%	98.614	50,33%
Diferença	2.259.078	2.467.061	356.817	394.783	-	53.228	-

Fonte: Própria autora

As estatísticas do *corpus* foram calculadas com a utilização do *software Fine Count*[13].

[13] Software produzido pela Tilti Systems versão 2.6.1942 cuja última versão 5 de dezembro de 2014. Disponível em: <http://www.tilti.com/software-for-translators/finecount/>. Acesso em: 10 de Janeiro 2017.

4.4 Clusterização dos Dados com o Modelo Cassiopeia

Na etapa de mineração, são aplicados métodos e algoritmos para a identificação de padrões interessantes e extração de conhecimento. Nessa etapa escolhe-se qual tarefa de acordo com a necessidade do usuário. A técnica de classificação escolhida para este trabalho foi a abordagem com *Cluster*ização, onde a necessidade foi verificar o grau de similaridade e a formação de grupos naturais de *tweets*.

Foram realizados testes para verificar qual a quantidade máxima de *tweets* era possível ser executada no Cassiopeia. Ao final dos testes a quantidade máxima de *tweets* suportada pelo modelo Cassiopeia foi de 700 *tweets*. Desse modo, o total de *tweets* processados foram 4.900. Obviamente, não é possível validar toda a análise com base só em 4.900 *tweets*, uma vez que foram coletados mais de 200.000. No próximo capítulo, serão analisados os resultados da análise de *clusters* e de outras informações provenientes dos *tweets* sobre o ENEM 2016.

A representação proporcional de cada amostra na nossa população é apresentada na Tabela 4:

Tabela 4 - Síntese de *Tweets* Coletados e Processados: *Corpus* ENEM 2016

Arquivos	Quantidade de *Tweet*	Porcentagem de *Tweet* Processados
***Corpus* Maio**		
Qtde	923	700
% do Total		76%
***Corpus* Agosto**		
Qtde	3.699	700
% do Total		19%
***Corpus* Setembro**		
Qtde	5.602	700
% do Total		12%
***Corpus* Outubro**		
Qtde	23.066	700
% do Total		3%
***Corpus* Novembro**		
Qtde	154.185	700
% do Total		0,45%
***Corpus* Dezembro**		
Qtde	36.916	700
% do Total		2%
***Corpus* Janeiro**		
Qtde	15.231	700
% do Total		5%
Corpus Total	239.622	4.900

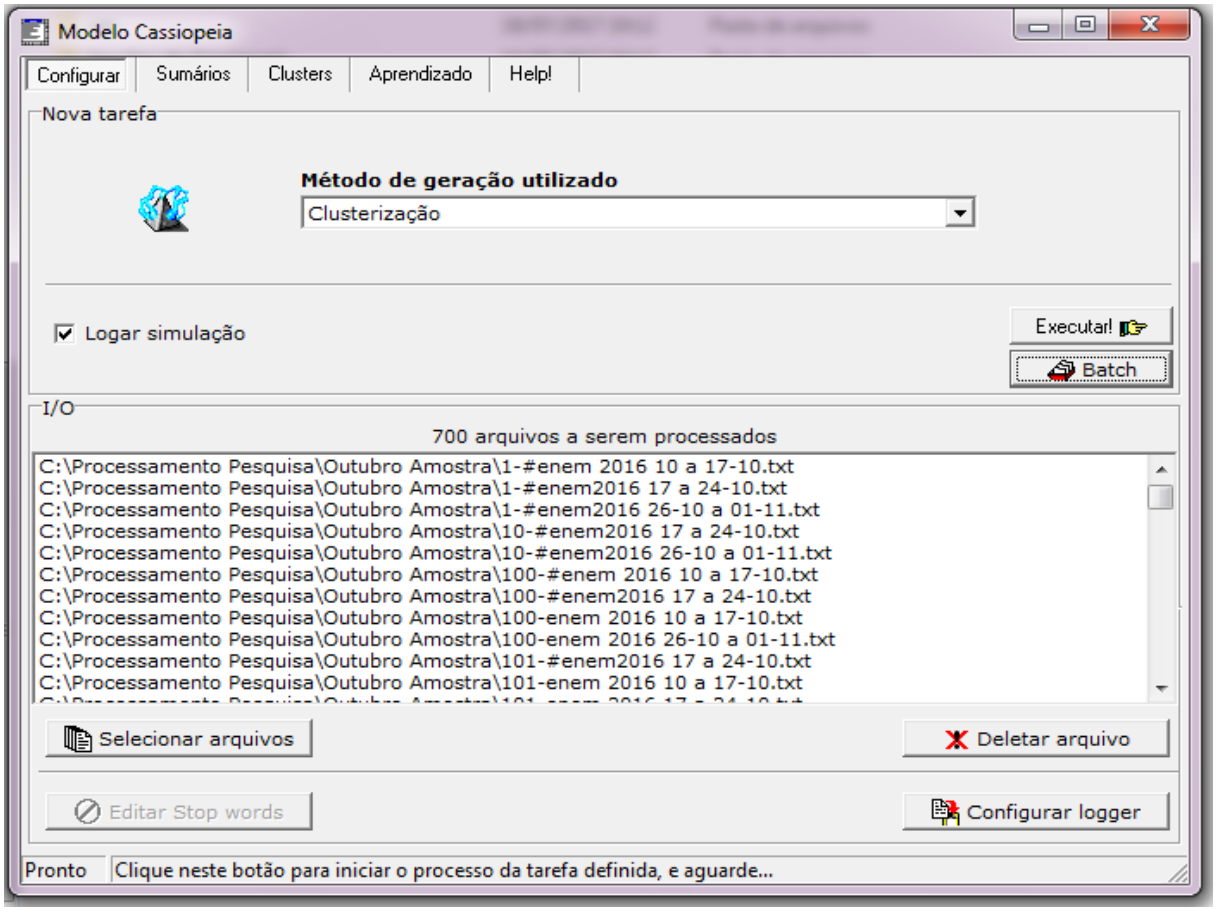
Fonte: Própria autora

Desta forma, foi possível definir a proporção de cada uma das 7 categorias na amostra a ser utilizada sobre o tema ENEM 2016 para identificação dos agrupamentos. A *Cluster*ização do *corpus* foi realizada utilizando 30 iterações, dessa forma, foram gerados resultados das médias finais acumuladas das medidas *Coesão* e *Acoplamento*.

Dessa maneira, foi realizada a *Cluster*ização nos 700 textos fontes, das 7 categorias do *corpus*. O resultado do processo de *Clusterização* são os *tweets* agrupados de acordo a similaridade identificada pelo algoritmo do modelo Cassiopeia e as métricas internas coesão, acoplamento e *coeficiente de silhouette* que mensura a qualidade dos agrupamentos. O Cassiopeia gerou uma média acumulada dessas sete categorias que se encontram nos gráficos dos experimentos.

A Figura 10 mostra a tela principal do modelo Cassiopeia. Modelo utilizado no processo de *Clusterização* para descobrir padrões e relacionamentos entre os textos.

Figura 10 - Tela Principal do Modelo Cassiopeia



Fonte: Própria autora

5 ANÁLISE DOS RESULTADOS

A fase de análise corresponde à interpretação dos resultados e podem ser mostrados de diversas formas, porém devem ser apresentadas de forma que o usuário possa entender e interpretar os resultados obtidos.

Como qualquer tarefa no campo da mineração de texto, os resultados sempre devemser encarados apenas como um indicador e não como verdade absoluta. O corpus utilizado é limitado e isso quer dizer que há palavras e expressões que podem não ter sido identificadas na etapa de clusterização dos *tweets*. Além disso, a linguagem falada na internet possui gírias, ironias e sarcasmo que não são reconhecidos em modelos simples de análise de texto. Os *tweets* também possuem muitos erros de digitação que dificultam a clusterização, pois o clusterizador também não reconhece essas palavras.

No caso deste trabalho, os *tweets* foram clusterizados, porém nem todos os *tweets* expressam opiniões significativas. Alguns deles apresentam apenas fatos ou comentários ambíguos que não cabem em nenhuma interpretação.

O resultado da pesquisa está dividido em duas partes: (1) apresentação dos resultados obtidos através do Cassiopeia, ou seja, os gráficos com as métricas utilizadas pelo mesmo; (2) visualização do conteúdo dos agrupamentos através de nuvem de palavras.

5.1 Resultados Obtidos Através do Cassiopeia

O *corpus* empregado na criação dos *clusters* foi o *corpus* de tema ENEM 2016 criado no trabalho de Silva e Guelpeli, (2017) que são compostos por 7 categorias: *Corpus* Maio, *Corpus* Agosto, *Corpus* Setembro, *Corpus* Outubro, *Corpus* Novembro, *Corpus* Dezembro *e Corpus* Janeiro.

A métrica adotada para mensurar a qualidade dos *clusters* gerados pelo modelo Cassiopeia foi a *Coeficiente de Silhouette*, devido sua importância, pois seu valor representa uma média harmônica entre as métricas acoplamento e coesão, combinando em uma única média as duas métricas.

Para Witten e Frank (2011), métodos estatísticos podem ser utilizados como forma de avaliação dos algoritmos, isto é, saber se o processo funcionou ou não como previsto.

Nesse caso, as métricas podem informar o percentual de similaridade entre os elementos do mesmo agrupamento e a similaridade média entre os agrupamentos para um determinado contexto.

O resultado do processo de mineração de textos pelo modelo Cassiopeia é mostrado através de gráficos. Os gráficos possuem os valores do eixo x correspondentes ao número de passos, ou número de vezes que cada *corpus* foi processado no algoritmo do Modelo Cassiopeia. Os eixos y apresentam os valores acumulados das métricas internas.

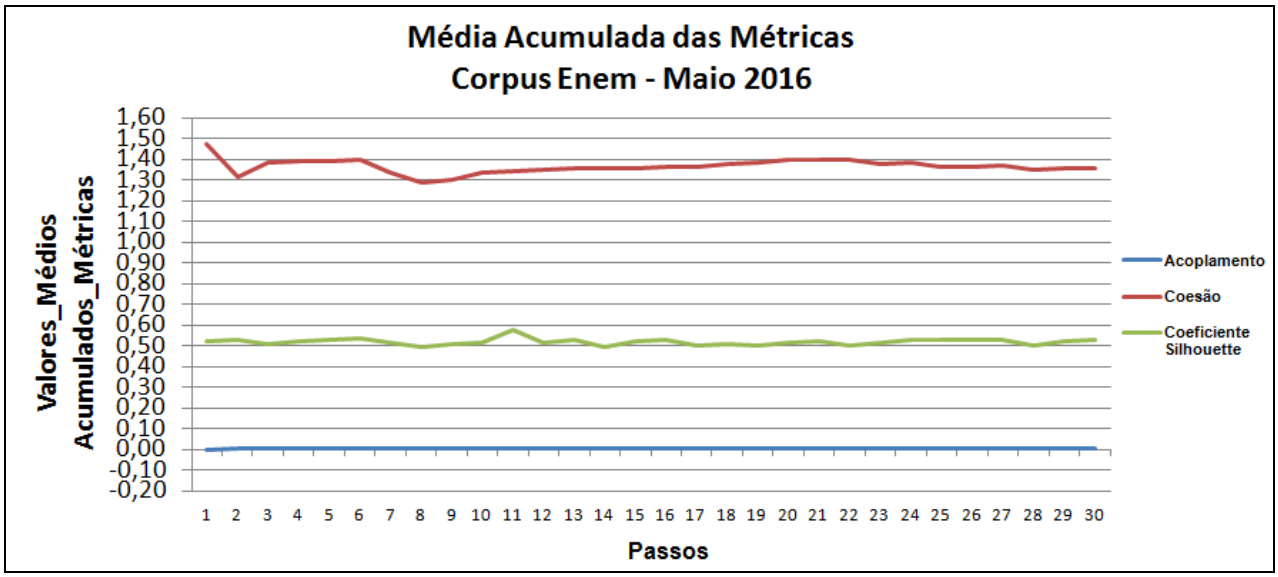
De acordo com a fundamentação teórica descrita no Capítulo 2, os valores de coesão consistem na similaridade entre documentos dentro de um mesmo *cluster*, então, quanto maior for esse valor de similaridade e quanto mais distante do 0 (zero), melhor é o resultado. *Cluster* de textos similares com valores de coesão mais próximos de 1(um) é um bom resultado, pois significa que os textos de cada *cluster* são altamente homogêneos e/ou similares. Os valores de acoplamento consistem na similaridade entre os *clusters* gerados, então, quanto menor for esse valor de similaridade e quanto mais próximo de 0 (zero), melhor é o resultado. *Clusters* com valores de acoplamento mais próximos de 0 (zero) é um bom resultado, pois significa que os textos de cada *cluster* são altamente heterogêneos e/ou diferentes dos outros *clusters*.

Coeficiente de *Silhouette* consiste na média harmônica entre o acoplamento e a coesão, ou seja, de quanto um objeto é similar aos demais membros do seu grupo, e de quanto este mesmo objeto é distante dos de outro grupo. Valores de coeficiente de *silhouette* mais próximos de 1 (um) é um bom resultado.

A análise de *cluster* busca agrupar elementos de textos baseando-se na similaridade entre eles, que consistem nos valores de coesão. Os grupos são determinados de forma a obter-se homogeneidade dentro dos grupos e heterogeneidade entre eles.

A Figura 11 apresenta o gráfico gerado através do modelo Cassiopeia com os resultados obtidos no processamento do mês de maio.

Figura 11 - Resultados obtidos pelo Modelo Cassiopeia *Corpus* ENEM 2016 do Mês de Maio



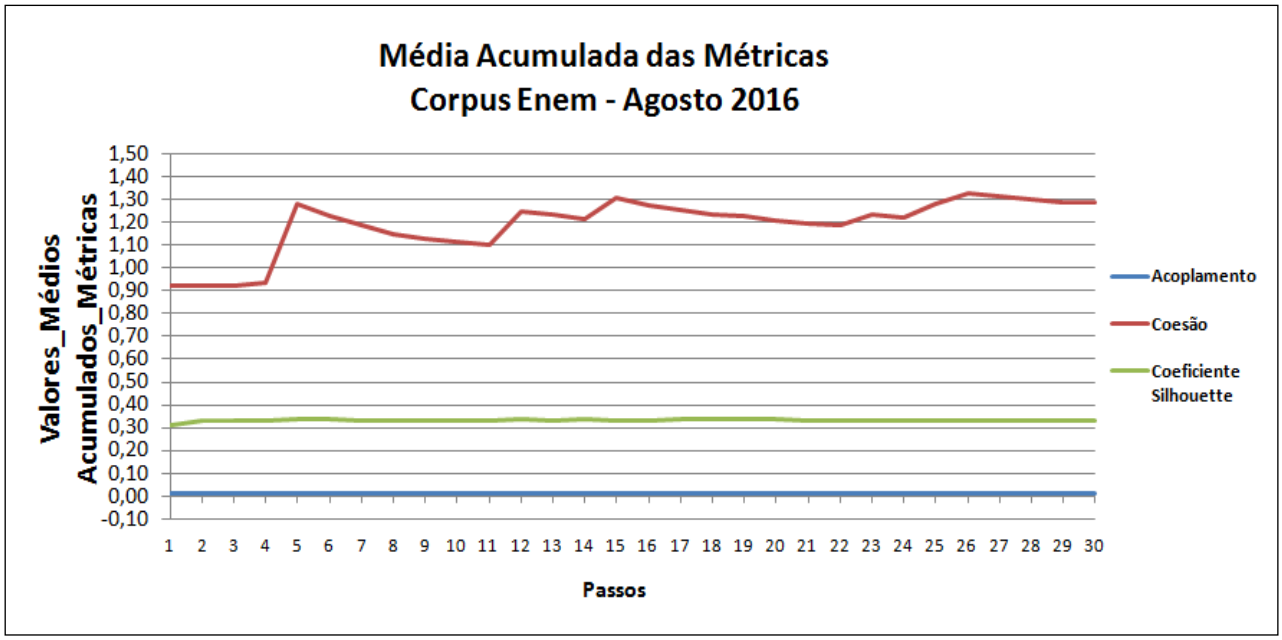
Fonte: Própria autora

No processamento do mês de Maio 2016 foram formados 322 *clusters*. O valor de coesão está entre 1,29 e 1,48 enquanto que o de acoplamento permaneceu em 0,00. O resultado do Coeficiente de *Silhouette* está entre 0,49 e 0,57. Os resultados foram satisfatórios considerando o baixo valor no acoplamento e altos valores na coesão. Os valores de coesão variam entre 0 e 1, e observa-se que neste processamento o Cassiopeia gerou valores de coesão acima de 1. Isso ocorreu devido o tamanho dos textos serem muito pequenos. O valor baixo de acoplamento obtido é considerado um valor bom, pois significa que os textos de um *cluster* são diferentes dos demais *clusters*, ou seja, o Cassiopeia realizou um bom agrupamento de textos.

Os resultados dos acoplamentos do *corpus* do ENEM do mês de Maio - 2016 mostraram bons resultados de acoplamento e coesão, visto que os textos dos agrupamentos são similares entre si e diferentes dos demais grupos.

A Figura 12 mostra os resultados dos testes no mês de Agosto 2016 usando o *corpus* de *tweets* sobre o ENEM.

Figura 12 - Resultados obtidos pelo Modelo Cassiopeia no Processamento do Corpus ENEM 2016 no Mês de Agosto 2016



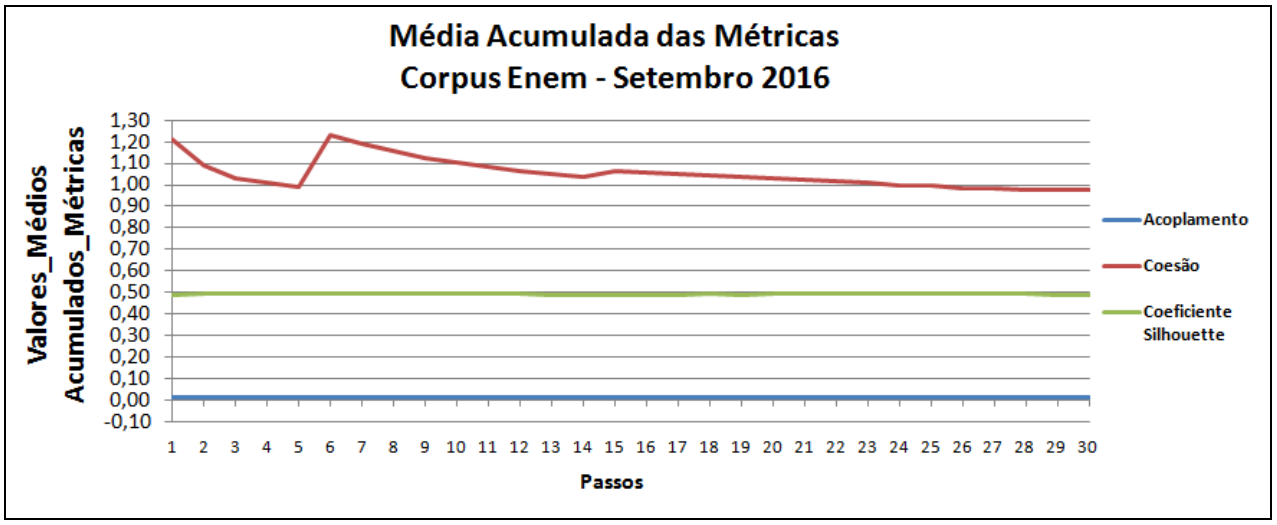
Fonte: Própria autora

No processamento desse mês o Cassiopeia gerou um total de 187 *clusters*. Os resultados da Figura 12 demonstram que os valores da média acumulada de coesão do *corpus*, apresentam valores entre 0,92 e 1,33. Os valores de coesão podem variar entre 0 e 1, e no processamento desse mês o Cassiopeia gerou o valor diferenciado 1,33. Isso ocorreu devido o tamanho dos textos serem muito pequenos. O valor de acoplamento permaneceu em 0,01, o que é considerado um valor bom, pois significa que os textos de um *cluster* são diferentes dos demais *clusters*, ou seja, o Cassiopeia realizou um bom agrupamento de textos. O resultado do Coeficiente de *Silhouette* no *corpus* de Agosto - ENEM 2016 está entre 0,31 e 0,34, que é a média dos valores de coesão e acoplamento.

Analisando os resultados dos acoplamentos do *corpus* de Agosto - ENEM 2016 percebe-se que o mesmo apresentou bons resultados de acoplamento e coesão, visto que os textos dos agrupamentos são altamente similares entre si e diferentes dos demais grupos.

A Figura 13 mostra os resultados dos testes no mês de Setembro 2016 usando o *corpus* de *tweets* sobre o ENEM.

Figura 13 - Resultados obtidos pelo Modelo Cassiopeia no Processamento do *Corpus* ENEM 2016 no Mês de Setembro 2016



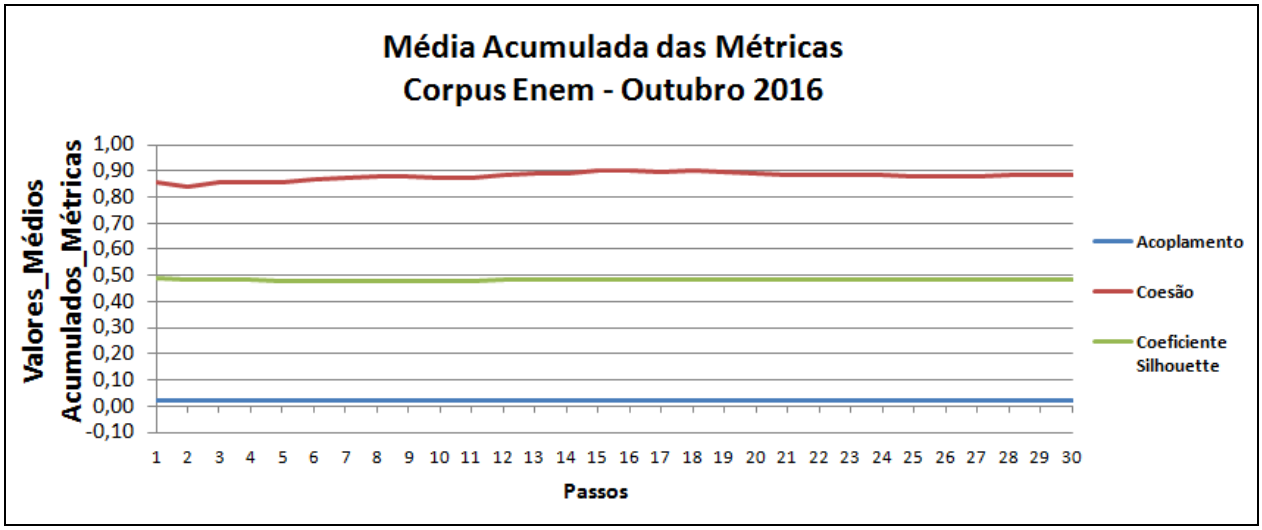
Fonte: Própria autora

No processamento do mês de setembro 2016 foram gerados 151 *clusters.* O valor de coesão está entre 0,97 e 1,21 enquanto que o de acoplamento 0,01. O resultado do Coeficiente de *Silhouette* está entre 0,48 e 0,49. Os resultados foram satisfatórios considerando o baixo valor no acoplamento e altos valores na coesão.

Diante dos resultados dos acoplamentos do *corpus* de Setembro - ENEM 2016 percebe-se que o mesmo apresentou bons resultados de acoplamento e coesão, visto que os textos dos agrupamentos são altamente similares entre si e diferentes dos demais grupos.

A Figura 14 mostra os resultados dos testes no mês de Outubro 2016 usando o *corpus* de *tweets* sobre o ENEM.

Figura 14 - Resultados obtidos pelo Modelo Cassiopeia no processamento do *Corpus* ENEM 2016 no Mês de Outubro 2016



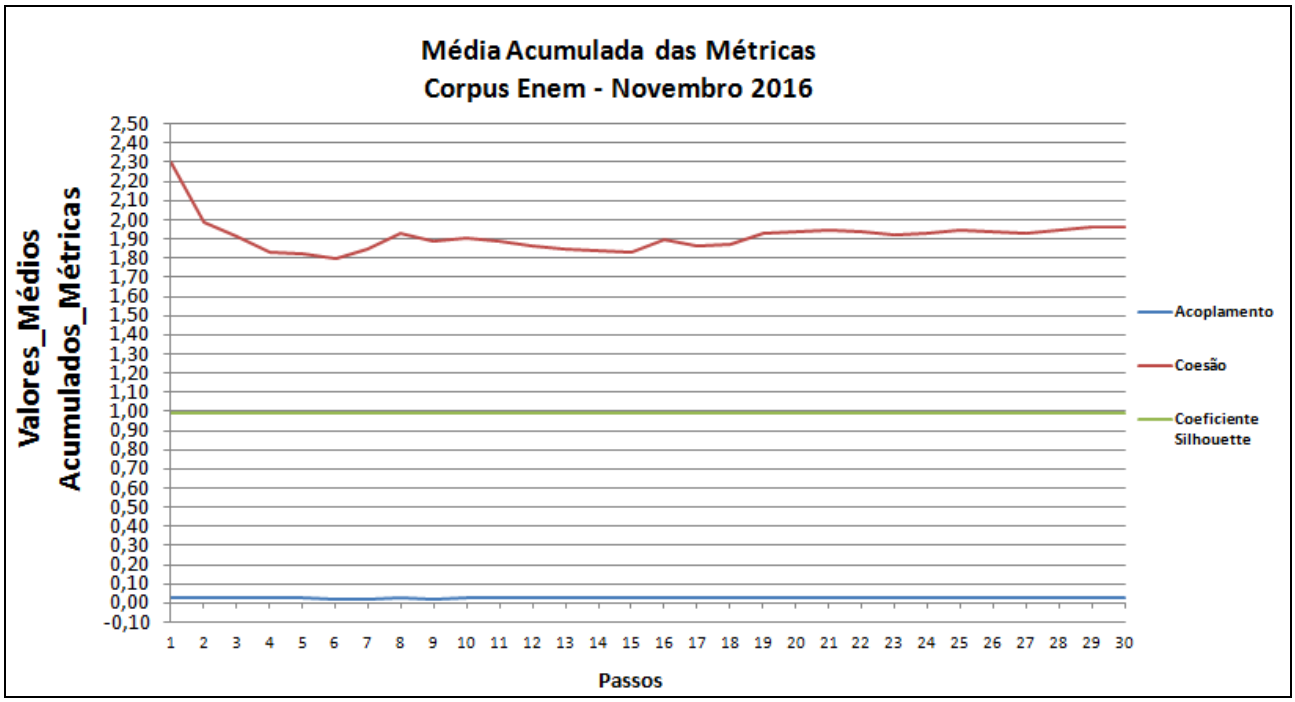
Fonte: Própria autora

No processamento do mês de outubro 2016 o Cassiopeia gerou um total de 100 *clusters*. O valor de coesão está entre 0,84 e 0,90 enquanto que o de acoplamento permaneceu em 0,02. O resultado do Coeficiente de *Silhouette* está entre 0,48 e 0,49.

Analisando os resultados dos acoplamentos do *corpus* de Outubro - ENEM 2016 percebe-se que o valor de coesão ficou abaixo de 1 mas ainda assim é considerado um bom resultado visto que os textos dos agrupamentos são similares entre si e diferentes dos demais grupos.

A Figura 15 mostra os resultados dos testes no mês de Novembro 2016 usando o *corpus* de *tweets* sobre o ENEM.

Figura 15 - Resultados obtidos pelo Modelo Cassiopeia no Processamento do *Corpus* do ENEM 2016 no Mês de Novembro 2016



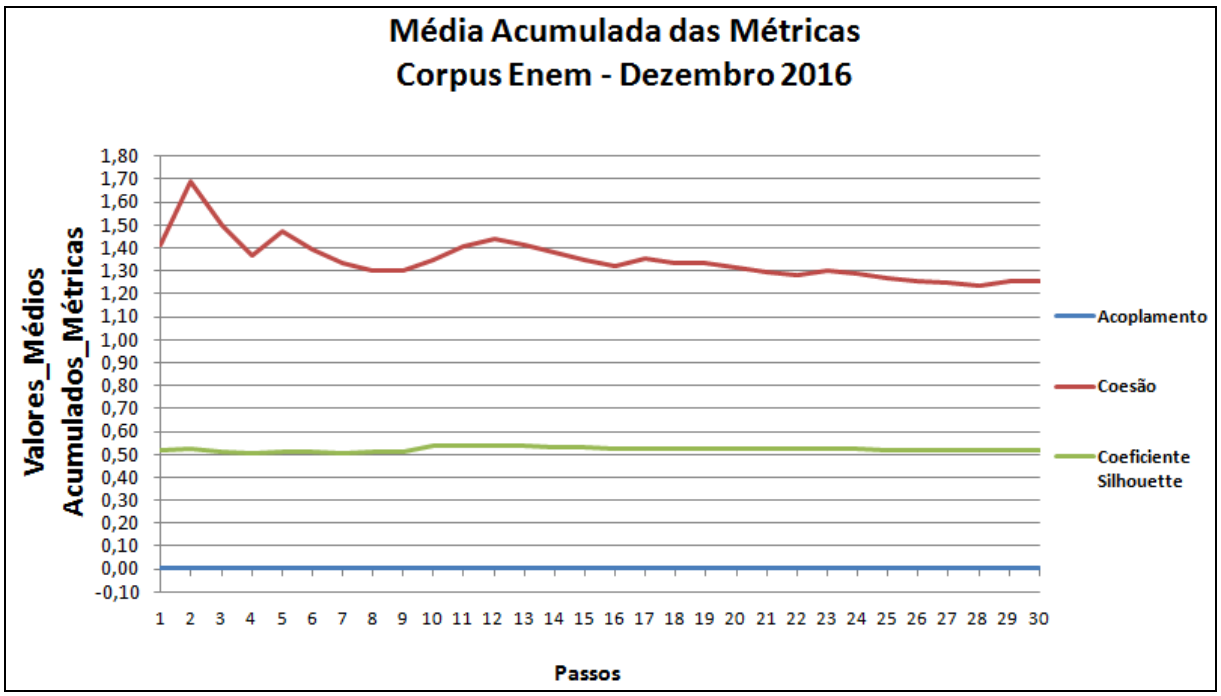
Fonte: Própria autora

No processamento do mês de Novembro 2016 o Cassiopeia gerou um total de 150 *clusters*. O valor de coesão está entre 1,80 e 2,30, ou seja, os textos dos agrupamentos são altamente similares entre si enquanto que o de acoplamento entre 0,02 e 0,03, ou seja, os agrupamentos são diferentes dos demais. O resultado do Coeficiente de *Silhouette* foi de 0,99.

O resultado dos acoplamentos do *corpus* de Novembro - ENEM 2016 foi satisfatório, pois agruparam os textos similares entre si e diferentes dos demais grupos.

A Figura 16 mostra os resultados dos testes no mês de Dezembro 2016 usando o *corpus* de *tweets* sobre o ENEM.

Figura 16 - Resultados obtidos pelo Modelo Cassiopeia no Processamento do *Corpus* ENEM 2016 no Mês de Dezembro 2016

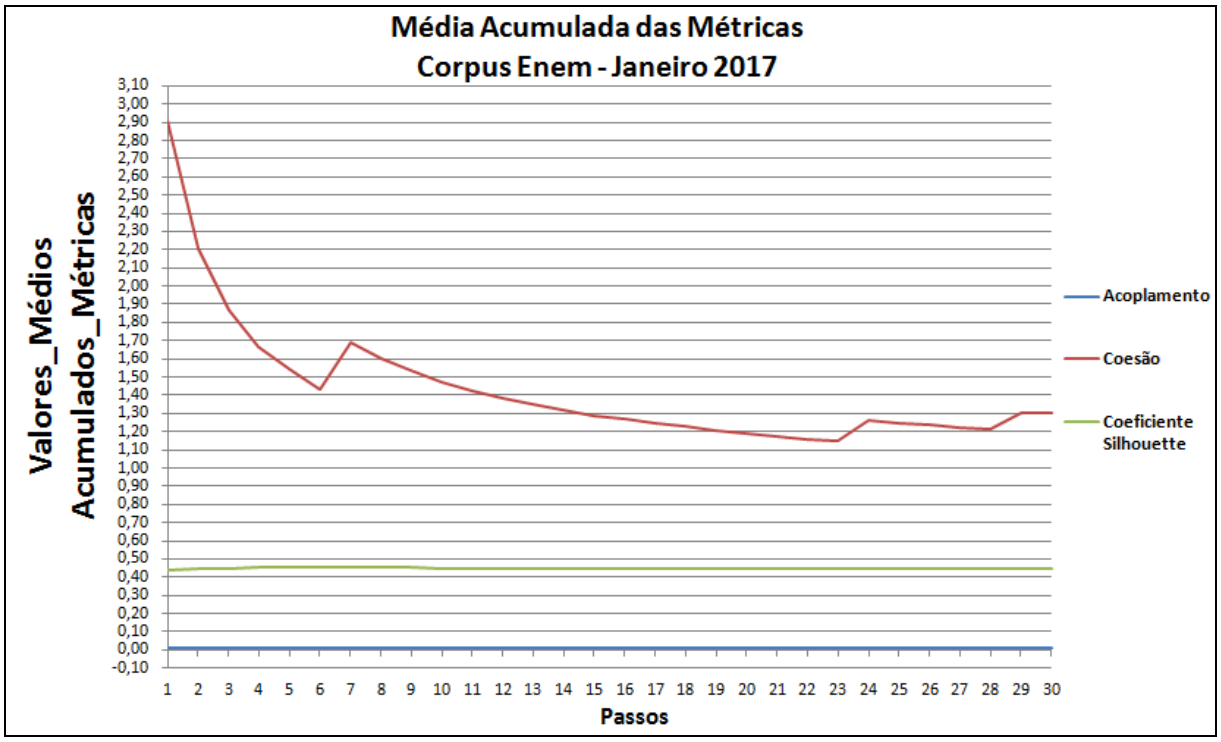


Fonte: Própria autora

No processamento do mês de Dezembro 2016 o Cassiopeia gerou um total de 290 *clusters*. O valor de coesão está entre 1,23 e 1,69 enquanto que o de acoplamento entá entre 0 e 0,01. O resultado do Coeficiente de *Silhouette* está entre 0,51 e 0,54. Assim, o resultado dos acoplamentos do *corpus* de Dezembro - ENEM 2016 foi satisfatório, pois os valores obtidos representam agrupamentos com textos similares entre si e diferentes dos demais grupos.

A Figura 17 mostra os resultados dos testes no mês de Janeiro 2016 usando o *corpus* de *tweets* sobre o ENEM.

Figura 17 - Resultados obtidos pelo Modelo Cassiopeia no Processamento do *Corpus* ENEM 2016 no Mês de Janeiro 2017



Fonte: Própria autora

No processamento do mês de janeiro 2017 o Cassiopeia gerou um total de 568 *clusters*. O valor de coesão está entre 1,15 e 2,91 enquanto que o de acoplamento permaneceu em 0,01. O resultado do Coeficiente de *Silhouette* está entre 0,44 e 0,45. Os resultados do processamento do *corpus* ENEM 2016 do mês de Janeiro 2017 demonstram que houve agrupamentos com bons resultados, pois representam *clusters* com textos similares entre si e diferentes entre os demais.

Pode-se concluir que o Modelo Cassiopeia atendeu aos propósitos de processamento fornecendo valores de acoplamento, coesão e coeficiente de silhouette satisfatórios em relação à quantidade de palavras e a qualidade dessas dentro de um *cluster*.

5.2 Visualização dos Agrupamentos Através da Nuvem de Palavras

Conforme comentado anteriormente, o corpus construído neste trabalho é composto por 7 categorias, que são os meses: maio, agosto, setembro, outubro, novembro, dezembro e janeiro. Foi realizado o processamento pelo Cassiopeia por categoria. Para ter uma melhor visualização dos resultados através da nuvem de palavras foi realizada uma seleção dos *clusters*. Como regra de corte, estabeleceu-se que seriam representados na nuvem de palavras somente os agrupamentos que obtiveram coesão maior ou igual à média de todos os valores de coesão obtidos no processamento, pois são mais representativos. A Tabela 5 mostra as categorias, a quantidade total de *tweets* coletados, quantidade utilizada nas amostras para realizar a mineração de texto, quantidade de *clusters* obtidos após o processamento no Cassiopeia, a quantidade de *tweets* aproveitados, pois alguns agupamentos possuíram valor de coesão 0 e não foram considerados, a média da coesão obtida no processamento de cada categoria e a quantidade dos *clusters* utilizados para a criação da nuvem de palavras, sendo que esses *clusters* foram os que obtiveram valores maiores ou iguais à média.

Tabela 5 – Síntese das Categorias de *Tweets*

Categorias	***Quantidade Total Tweets***	***Quantidades Tweets Utilizados em Amostra***	***Quantidade Total Clusters***	***Quantidade Clusters Aproveitados***	***Média/Coesão Clusters***	***Quantidade Clusters Utilizados em Nuvem***
Maio	923	700	322	169	12,40	59
Agosto	3.699	700	187	187	12,53	74
Setembro	5.602	700	151	145	13,64	60
Outubro	23.066	700	100	95	14,13	44
Novembro	154.185	700	150	148	21,60	43
Dezembro	36.916	700	290	104	16,71	51
Janeiro	15.231	700	568	49	10,92	20

Fonte: própria autora

Em seguida, serão apresentados, separadamente, os resultados das análises referentes a cada uma das categorias do trabalho. Serão apresentadas também, algumas correlações encontradas nos agrupamentos.

5.2.1 Análise Mês Maio 2016

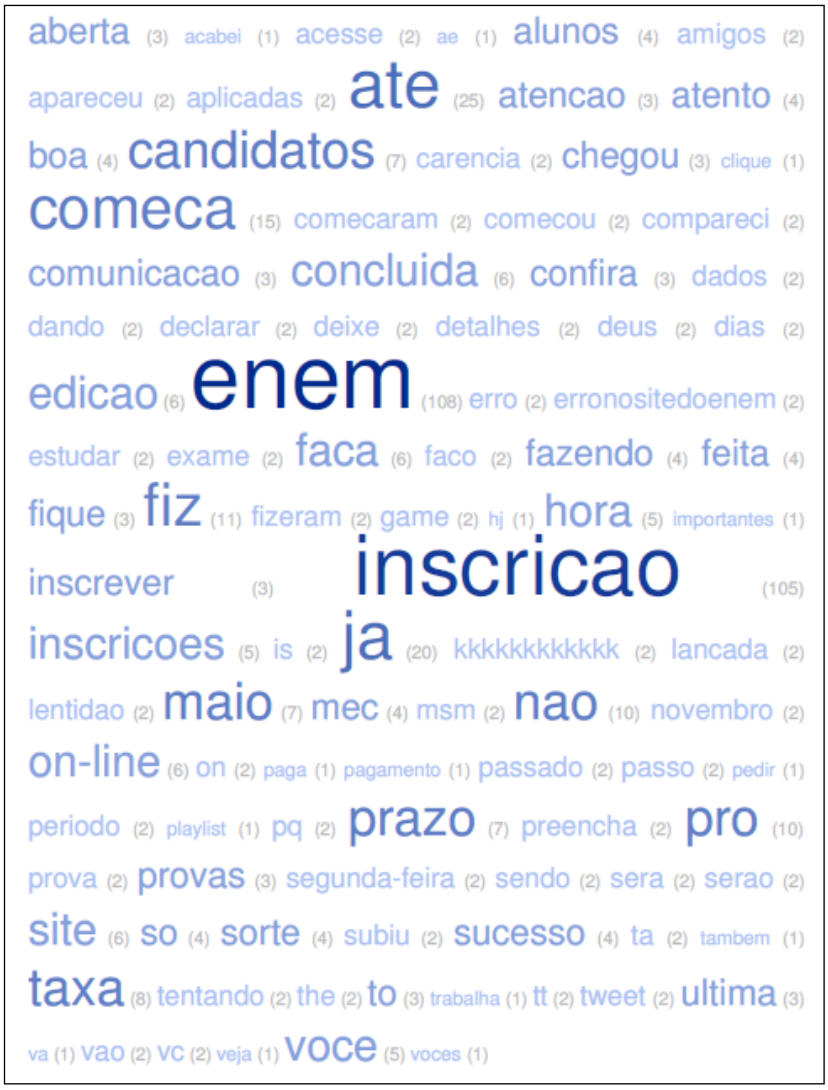
A Tabela 6 apresenta as palavras mais utilizadas em *Tweets* no mês de Maio 2016. Em seguida, a Figura 19 apresenta a nuvem de palavras obtida através dos *tweets* presentes nos *clusters* com coesão maior ou igual à média no mês de Maio 2016:

Tabela 6- Frequência das palavras mais utilizadas em *Tweets* Maio 2016

Palavra	***Frequência***
ENEM	108
INSCRIÇÃO	105
ATÉ	25
JÁ	20
COMEÇA	15
FIZ	11
TAXA	8
PRAZO	7
CANDIDATO	7
FAÇA	6
CONCLUÍDA	6

Fonte: própria autora

Figura 18- Nuvem de Palavras Maio 2016



Fonte: Própria autora

Analisando a Tabela 6 e a Figura 18, é possível perceber a predominância de *twests* que expressam a preocupação com o período de inscrições, expressa pelas palavras "começa" e "até" e se já concluíram ou não as inscrições do ENEM por meio das palavras "concluída" e "fiz". As Figuras 19 e 20 mostram exemplos desses *tweets*:

Figura 19- *Tweet* de *Cluster* Enem Maio 2016

As inscricoes do Enem comecaram esta semana e vao ate o dia 20 de Maio as 23h59.

Fonte: Própria autora

Figura 20 - *Tweet* de Cluster Enem Maio 2016

Minha inscricao para o ENEM 2016, ja foi concluida com sucesso. Nao deixe para fazer a sua de ultima hora.

Fonte: Própria autora

Também é possível observar *tweets* relacionados à "aumento da taxa de inscrição do ENEM" por meio das palavras "subiu"e "taxa" e relacionados a "lentidão no site de inscrições" por meio da palavra "lentidão". Os *tweets* das Figuras 21 e 22 são exemplos sobre o assunto em questão:

Figura 21 - *Tweet* de *Cluster* Enem Maio 2016

Inscricoes para o Enem 2016 vao ate o dia 20 de maio: Este ano houve um aumento na taxa de inscricao, que pas...

Fonte: Própria autora

Figura 22 - *Tweet* de *Cluster* Enem Maio 2016

Taxa, lentidao e erro: alunos reagem a abertura de inscricao do #Enem2016 #G1

Fonte: Própria autora

5.2.2 Análise Mês Agosto 2016

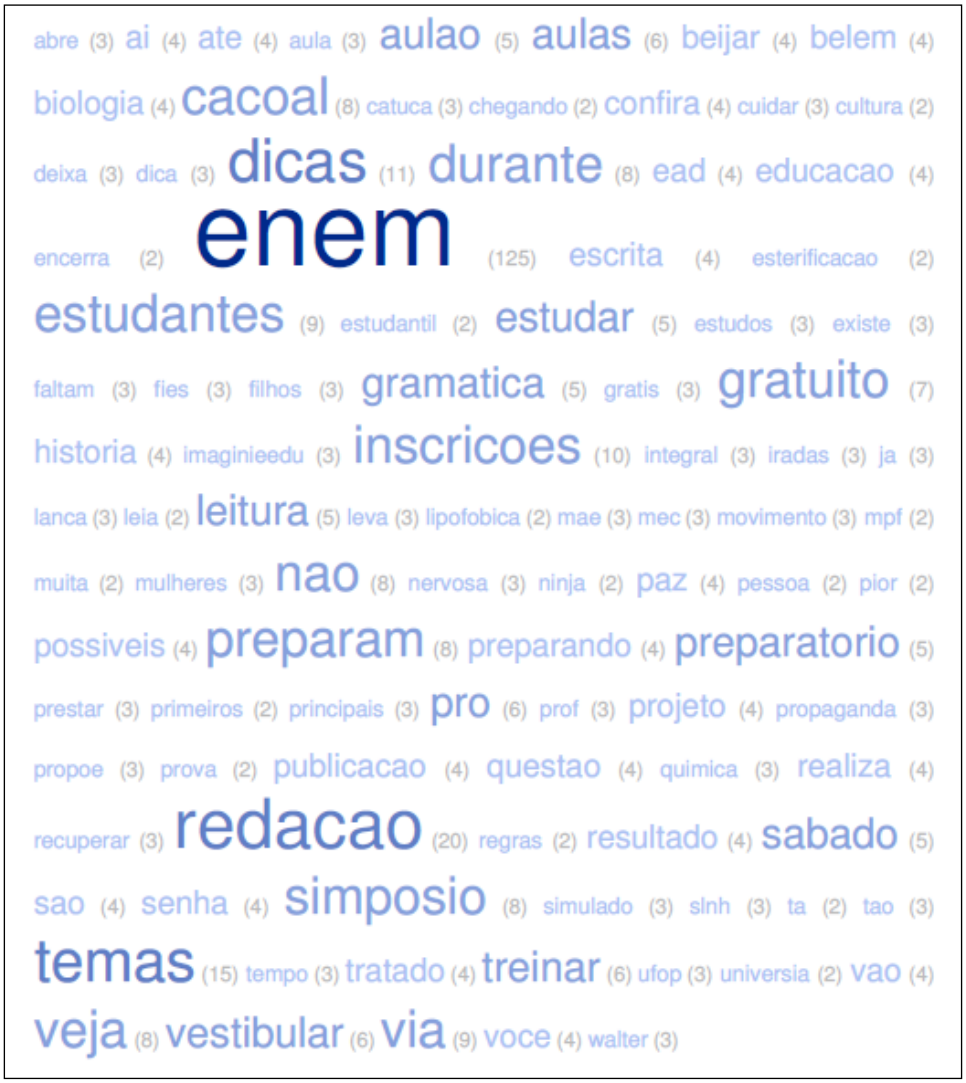
A Tabela 7 apresenta as palavras mais utilizadas em *Tweets* no mês de Agosto 2016. Em seguida, a Figura 23 apresenta a nuvem de palavras obtida através dos *tweets* presentes nos *clusters* com coesão maior ou igual à média no mês de Agosto 2016:

Tabela 7 - Frequência das palavras mais utilizadas em *Tweets* Agosto 2016

Palavras	***Frequência***
ENEM	125
REDAÇÃO	20
TEMAS	15
DICAS	11
INCRIÇÕES	10
ESTUDANTES	9
PREPARAM	8
SIMPÓSIO	8
PREPARATÓRIO	5

Fonte: Própria autora

Figura 23 - Nuvem de Palavras Agosto 2016

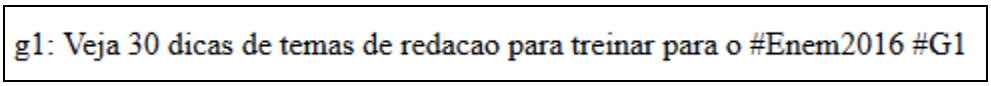


Fonte: Própria autora

A Tabela 7 e a Figura 23 permitem perceber a predominância de *tweets* que expressam a preparação para a prova do ENEM 2016 e que fazem suposições de "temas para a redação", uma vez que a prova foi aplicada no mês de novembro e o mês de agosto antecede a realização das provas. Assim, os alunos no mês de Agosto estão em período de preparação.

Essas observações são confirmadas pelo aparecimento das palavras "dicas", "temas" e "redação". A Figura 24 é um exemplo de *tweet* desse *cluster*:

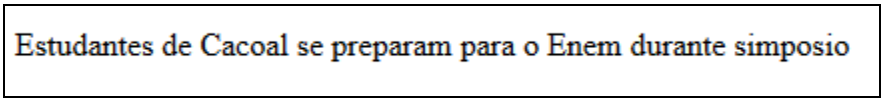
Figura 24 - *Tweet* de *Cluster* Enem Agosto 2016



Fonte: Própria autora

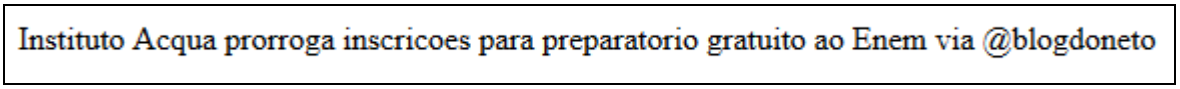
Também é possível ver que houve *tweets* reacionados à oferta de cursos preparatórios para o exame e alertas de inscrição para esses cursos através das palavras “preparam”, “preparatório”, “inscrições” e “simpósio”. Isso se confirma através dos exemplos das Figuras 25 e 26.

Figura 25- *Tweet* de *Cluster* Enem Agosto 2016



Fonte: Própria autora

Figura 26 - *Tweet* de *Cluster* Enem Agosto 2016



Fonte: Própria autora

5.2.3 Análise Mês Setembro 2016

A Tabela 8 apresenta as palavras mais utilizadas em *Tweets* no mês de Setembro 2016. Em seguida, a Figura 27 apresenta a nuvem de palavras obtida através dos *tweets* presentes nos *clusters* com coesão maior ou igual à média no mês de Setembro 2016:

Tabela 8- Frequência das palavras mais utilizadas em *Tweets* Setembro 2016

Palavras	***Frequência***
ENEM	103
SIMULADO	18
DICAS	14
HORA	13
MATEMÁTICA	12
TERMINA	10
PRAZO	10
INSCRIÇÕES	9
VIDEOS	9
ESTUDAR	9
HISTÓRIA	8

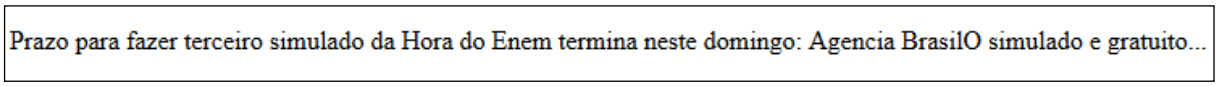
Fonte: Própria autora

Figura 27- Nuvem de Palavras Setembro 2016

Fonte: Própria autora

Analisando a Tabela 8 e a Figura 27, é possível perceber a predominância do uso de palavras que expressam a atenção ao prazo de realização do terceiro simulado da Hora do Enem através das palavras “simulado”, “hora”, “termina” e “prazo”. Lembrando que o MEC tornou disponível a terceira etapa do simulado de 2016 no portal Hora do Enem no período de 03 a 11 de setembro. A Figura 28 é um exemplo de *tweet* sobre desse *cluster*.

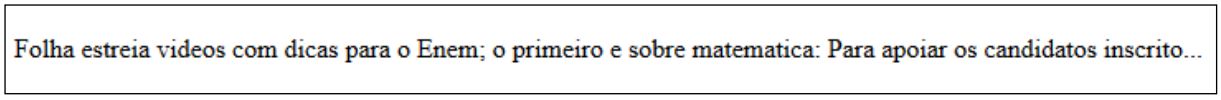
Figura 28 - *Tweet* de *Cluster* Enem Setembro 2016



Fonte: Própria autora

Também é possível observar palavras relacionadas à *links* de jornais onde são disponibilizados materiais como vídeos, artes, textos com dúvidas sobre a prova e apostas de estudos sobre cada disciplina, tais como "dicas", "matemática", "vídeos" e "história". O que pode ser visto com o *twitter* da Figura 29:

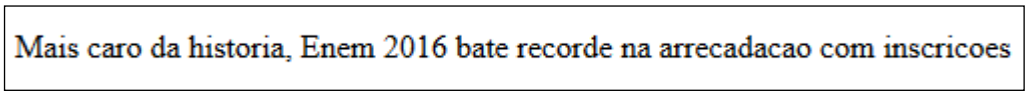
Figura 29 - *Tweet* de *Cluster* Enem Setembro 2016



Fonte: Própria autora

Outro assunto comentado foi a respeito do recorde na arrecadação com as inscrições do Enem 2016, considerado o mais caro da história através das palavras "inscrições", "recorde" e "história". Segue exemplo de *tweet* desse *cluster* com a Figura 30:

Figura 30- *Tweet* de *Cluster* Enem Setembro 2016



Fonte: Própria autora

5.2.4 Análise Mês Outubro 2016

A Tabela 9 apresenta as palavras mais utilizadas em *Tweets* no mês de Outubro 2016. Em seguida, a Figura 31 apresenta a nuvem de palavras obtida através dos *tweets* presentes nos *clusters* com coesão maior ou igual à média no mês de Outubro 2016:

Tabela 9 - Frequência das palavras mais utilizadas em *Tweets* Outubro 2016

Palavras	***Frequência***
ENEM	82
ESCOLAS	32
SIMULADO	25
INEP	14
IFAP	14
LISTA	14
NOTAS	14
PUBLICAS	14
BRASIL	13
FORA	13

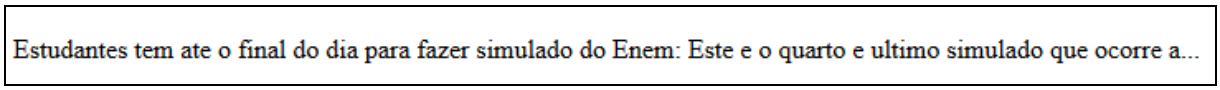
Fonte: Própria autora

Figura 31- Nuvem de Palavras Outubro 2016

Fonte: Própria autora

Analisando a Tabela 9 e a Figura 31 em conjunto, é possível perceber que houve *tweets* sobre a realização do 4º Simulado que foi realizado entre os dias 08 e 23 de outubro de 2016, através da palavra “simulado”. Na Figura 32 a seguir temos um exemplo desse assunto no *cluster* em questão:

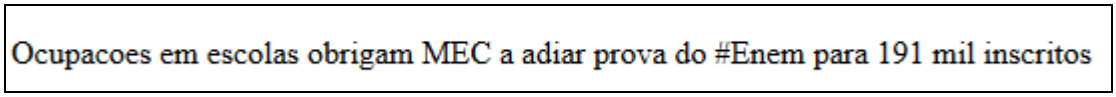
Figura 32 - *Tweet* de *Cluster* Enem Outubro 2016



Fonte: Própria autora

Também foi possível perceber *tweets* sobre o movimento de ocupação das escolas ocorrido em 2016 e sobre o adiamento do Enem em várias escolas do país, através das palavras “escolas”, “adiado” e “ocupações”. A Figura 33 é um exemplo de *tweet* sobre o assunto:

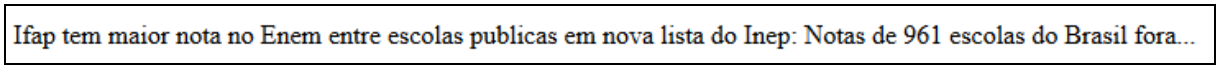
Figura 33 - *Tweet* de *Cluster* Enem Outubro 2016



Fonte: Própria autora

Outro assunto encontrado nos *tweets* dos *clusters* de outubro 2016 foi a divulgação da lista das notas médias obtidas pelas escolas na edição do Enem de 2015. O Instituto Federal do Amapá (Ifap) - Campus Macapá, apareceu como líder com a maior nota entre as escolas públicas do estado, e foi assunto comentado pelos usuários do *twitter* confirmado pela predominância das palavras “Inep”, “Ifap”, “lista” e “notas”. A Figura 34 mostra um exemplo de *tweet* desse assunto:

Figura 34 - *Tweet* de *Cluster* Enem Outubro 2016



Fonte: Própria autora

5.2.5 Análise Mês Novembro 2016

A Tabela 10 apresenta as palavras mais utilizadas em *Tweets* no mês de Novembro 2016. Em seguida, a Figura 35 apresenta a nuvem de palavras obtida através dos *tweets* presentes nos *clusters* com coesão maior ou igual à média no mês de Novembro 2016:

Tabela 10 - Frequência das palavras mais utilizadas em *Tweets* Novembro 2016

Palavras	***Frequência***
ENEM	522
TREINEIROS	110
ISENÇÃO	109
CANDIDATO	106
MEC	105
RETROCESSO	88
YOUTUBE	73

Fonte: Própria autora

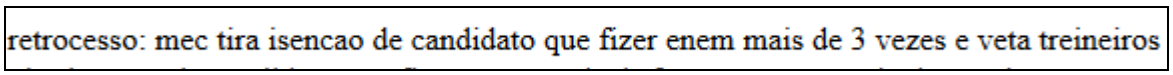
Figura 35- Nuvem de Palavras Novembro 2016

Fonte: Própria autora

Analisando a Tabela 10 e a Figura 35, é possível constatar *tweets* relacionados à medida elaborada pelo MEC para diminuir os custos da prova e anunciada no mês de novembro 2016, através das palavras “treineiros”, “isenção”, “candidato” e “retrocesso”. A medida prevê o cancelamento da adesão de candidatos treineiros, o limite de isenção por até

3 edições a cada candidato e deixa de servir para obtenção do certificado de conclusão do ensino médio. Na Figura 36 a seguir temos um exemplo desse assunto no *cluster* em questão:

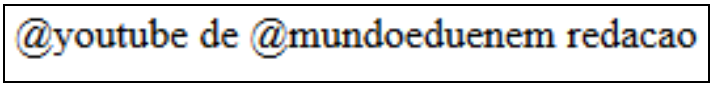
Figura 36 - *Tweet* de *Cluster* Enem Novembro 2016



Fonte: Própria autora

Também é possível perceber através da predominância da palavra "youtube" mostrada na Figura 37 que vídeos sobre o Enem são vistos no youtube e seus links são compartilhados na rede social *twitter*.

Figura 37 - *Tweet* de *Cluster* Enem Novembro 2016



Fonte: Própria autora

5.2.5 Análise Mês Dezembro 2016

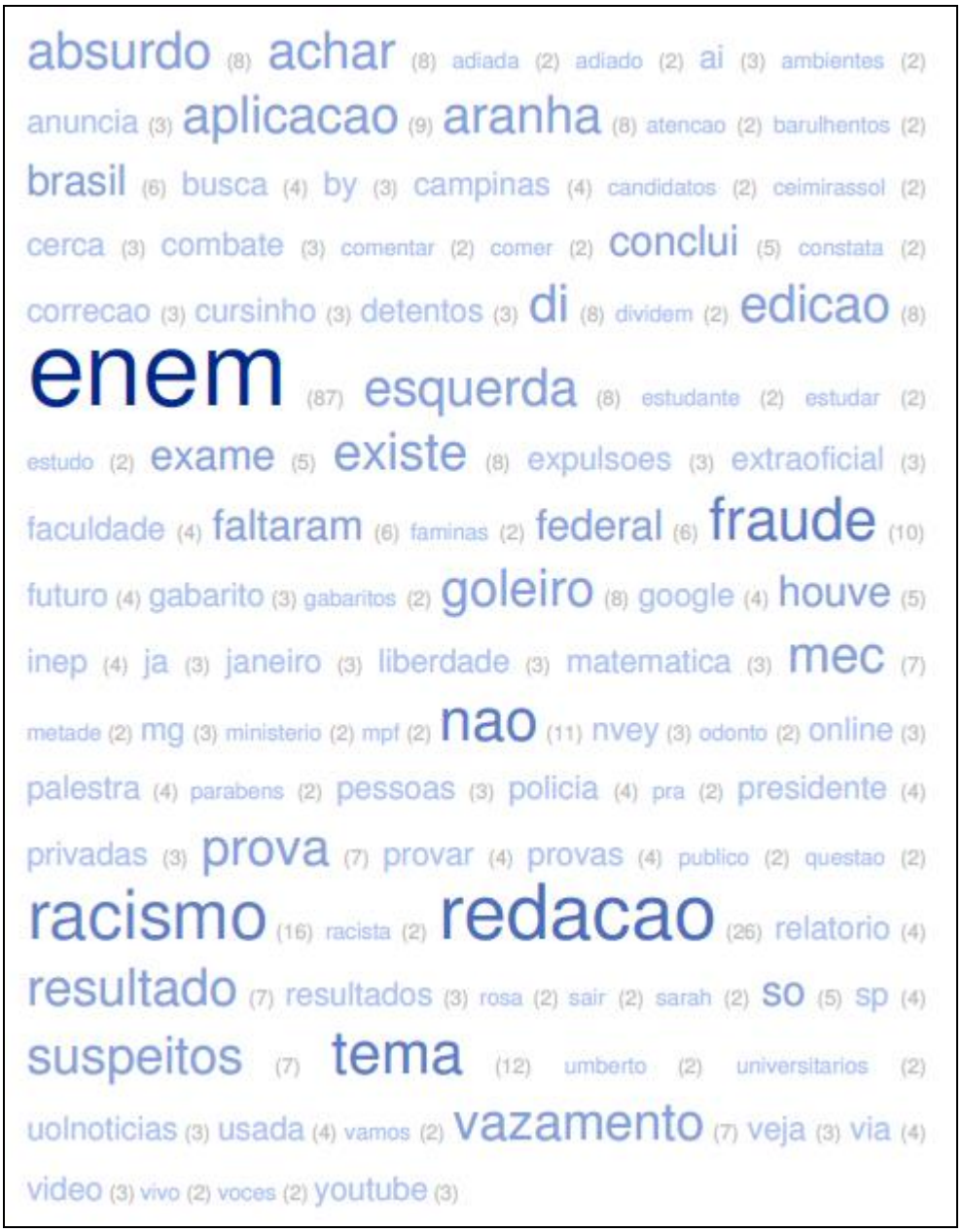
A Tabela 11 apresenta as palavras mais utilizadas em *Tweets* no mês de Dezembro 2016. Em seguida, a Figura 38 apresenta a nuvem de palavras obtida através dos *tweets* presentes nos *clusters* com coesão maior ou igual à média no mês de Dezembro 2016:

Tabela 11 - Frequência das palavras mais utilizadas em *Tweets* Dezembro 2016

Palavras	***Frequência***
ENEM	87
REDAÇÃO	26
RACISMO	16
TEMA	12
APLICAÇÃO	9
ARANHA	8
GOLEIRO	8
VAZAMENTO	7

Fonte: Própria autora

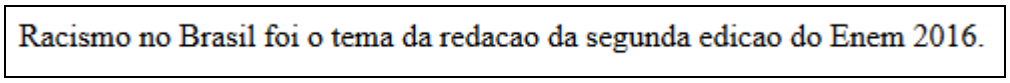
Figura 38- Nuvem de Palavras Dezembro 2016



Fonte: Própria autora

A Tabela 11 e a Figura 38 demonstram a predominância das palavras “Enem”, “redação”, “racismo”, “aplicação” e “tema”. São comentários que apontam o tema da redação da segunda aplicação do Enem 2016 "Caminhos para combater o racismo no Brasil". A Figura 39 mostra um exemplo de *tweet* desse *cluster*.

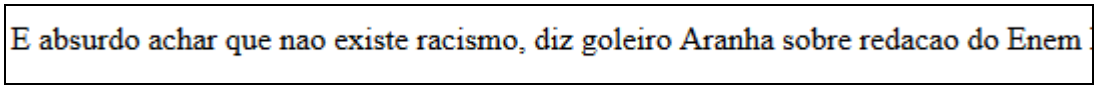
Figura 39 - *Tweet* de *Cluster* Enem Dezembro 2016



Fonte: Própria autora

Também é possível perceber através da predominância das palavras “aranha” e “goleiro” que o comentário do goleiro Aranha sobre o tema da redação do Enem 2016 "Caminhos para combater o racismo no Brasil" foi mencionado. O goleiro sofreu ofensas racistas proferidas por torcedores do time rival, durante uma partida de futebol. A Figura 40 mostra um exemplo de *tweet* sobre o assunto.

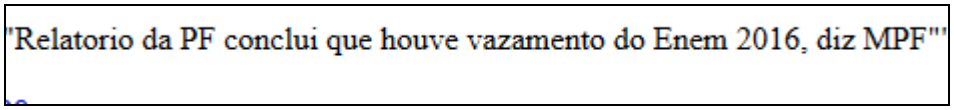
Figura 40 - *Tweet* de *Cluster* Enem Dezembro 2016



Fonte: Própria autora

Outro evento apontado através da predominância da palavra “vazamento” foi a ocorrência de vazamento das provas do Enem 2016. O *tweet* da Figura 41 é um exemplo apresentado.

Figura 41 - *Tweet* de *Cluster* Enem Dezembro 2016



Fonte: Própria autora

5.2.6 Análise Mês Janeiro 2017

Por fim, na última categoria temos a Tabela 12 com as palavras mais utilizadas em *Tweets* no mês de Janeiro 2016. Em seguida, a Figura 42 apresenta a nuvem de palavras obtida através dos *tweets* presentes nos *clusters* com coesão maior ou igual à média no mês de Janeiro 2017:

Tabela 12 - Frequência das palavras mais utilizadas em *Tweets* Janeiro 2017

Palavras	***Frequência***
ENEM	44
NOTA	18
CANDIDATOS	12
CONSEGUIRAM	12
MILHÕES	12
RESULTADO	7
NOTAS	6

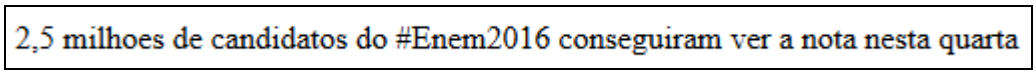
Fonte: Própria autora

Figura 42- Nuvem de Palavras Janeiro 2017

Fonte: Própria autora

A Tabela 12 e a Figura 42 demonstram a predominância das palavras “Enem”, “candidatos”, “conseguiram”, “milhões”, “nota”, “notas” e “resultado”. Todas as palavras estão relacionadas á divulgação dos resultados do Enem 2016. A Figura 43 apresenta um exemplo de *tweet* dos *clusters* formados no processamento do mês de Janeiro 2017:

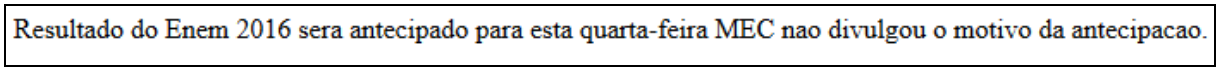
Figura 43 - *Tweet* de *Cluster* Enem Janeiro 2017



Fonte: Própria autora

Também foram encontrados *tweets* relatando antecipação de resultado do Enem 2016 através da predominância da palavra "resultado", como pode ser comprovado no *tweet* mostrado na Figura 44:

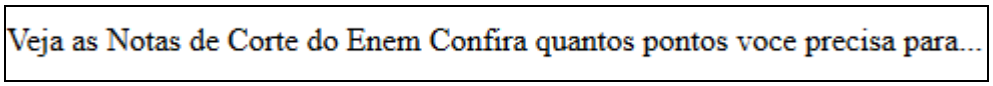
Figura 44 - *Tweet* de *Cluster* Enem Janeiro 2017



Fonte: Própria autora

O evento notas de corte foi apontado através da predominância da palavra "notas". Após a divulgação da nota do Enem os candidatos passam a analisar quais as notas de corte dos cursos e universidades pretendidos. O *tweet* da Figura 45 é um exemplo desse assunto.

Figura 45 - *Tweet* de *Cluster* Enem Janeiro 2017



Fonte: Própria autora

6 CONCLUSÃO

Redes sociais online se tornaram extremamente popular e é um espaço onde grandes quantidades de conteúdo são compartilhadas, e milhões de usuários interagem através de elos sociais de maneira espontânea. Com a mineração de textos nas redes sociais, surge uma série de oportunidades a serem exploradas, onde podemos identificar e prever comportamentos de grupos de pessoas.

Atualmente, a mineração de texto tornaram-se tarefas imprescindíveis para as empresas que querem conhecer o que os seus clientes estão falando da sua marca nas redes sociais, ao mesmo tempo em que sejam capazes de extrair informações que os possibilitem traçar estratégias competitivas que auxilie em seus negócios.

Ao realizar a análise dos comentários sobre o ENEM no *twitter* fica claro a explosão de conteúdo que está disponível e que as funções disponíveis e que a mineração de textos facilita de maneira crucial o entendimento desses textos. A mineração de textos com clusterização demonstrou ser uma maneira poderosa para resumir grandes volumes de textos e identificar palavras-chave ou objetos que facilitam o entendimento do que foi demonstrad.

Conforme afirmou (Russell, 2013) a partir de um tema arbitrário de interesse, o poder do *Twitter* e o conhecimento que se pode ganhar minerando seus dados textuais tornam-se muito mais evidente.

Este trabalho apresentou um processo de mineração de texto com clusterização dos comentários coletados do *Twitter* sobre o "ENEM 2016". Os *tweets* foram coletados do dia 09 de Maio de 2016 (inscrições do ENEM 2016) até o dia de divulgação dos resultados (18 de janeiro de 2017), para que assim, fosse feita uma análise dos acontecimentos relacionados ao evento durante todo o seu período de realização. Mais de 200.000 *tweets* foram coletados. Os *tweets*, então, foram clusterizados com a aplicação do Modelo Cassiopeia (Guelpeli, 2012).

Apesar dos bons resultados de acoplamento e coesão promovidos pela etapa de mineração, não é possível validar toda a análise com base só em 4.900 *tweets*, uma vez que foram coletados mais de 200.000. O motivo está na limitação do Modelo Cassiopeia utilizado no processamento.

A partir das análises, pode-se dizer que o Modelo Cassiopeia apresentou valores pertinentes quanto a qualidade de palavras dos *clusters* e que o processo de mineração de textos apresentou algumas informações significativas, mas que poderia ter gerado melhores resultados se fosse clusterizado todo o corpus. Os experimentos realizados mostraram que foi válido minerar os *tweets* obtidos da rede social *Twitter*, pois a técnica apresenta-se como uma boa alternativa para solucionar problemas relacionados á análise de dados textuais, porém o Modelo Cassiopeia possui limitação de processamento e precisa ser melhorado para que os resultados sejam mais bem sucedido.

Segundo a análise dos resultados, foi possível elencar alguns eventos significativos:

- Aumento da taxa de inscrição do ENEM e consequente recorde na arrecadação com as inscrições do ENEM 2016, considerado o mais caro da história;
- Identificação de novas fontes relacionadas à preparação para o ENEM como o acesso a vídeos no *youtube* sobre o exame, e compartilhamento dos links no *Twitter*;
- Aplicação de simulados através do Hora do ENEM;
- Movimento de ocupação das escolas ocorrido em 2016 e adiamento do ENEM em várias escolas do país;
- Medida que prevê o cancelamento da adesão de candidatos treineiros;
- Limite de isenção da taxa de inscrição do ENEM por até 3 edições a cada candidato;
- Medida que cria nova prova para obtenção do certificado de conclusão do ensino médio e que retifica os objetivos do ENEM.

Apesar das dificuldades, os resultados finais obtidos confirmam que a utilização da Mineração de Texto torna-se viável para processar o conteúdo gerado pelos usuários na rede social *Twitter* sobre o tema ENEM 2016. O Cassiopeia apresentou resultados coesos, o que mostra que o modelo é uma solução viável para *cluster*izar um volume de até 700 *tweets* por processamento e gerar *clusters* ou grupos de textos para serem analisados.

A respeito da rede social *Twitter* evidenciou-se através da mineração que existe a interação, por meio da conexão, onde várias pessoas no mundo todo expõem suas opiniões, sejam elas conhecidas ou não.

Quanto à mineração de textos oriundos de redes sociais demonstra ser uma prática importante para a análise e descoberta de tendências, sendo um procedimento ágil para descoberta de conhecimento. A mineração de dados textuais torna-se, então, uma das maiores fontes de investigação de informações e que além de ajudar na presente pesquisa, pode também ajudar em outras.

O campo de mineração de texto ainda está em processo de evolução e, como comentado no capítulo anterior, os resultados devem ser tratados como indicadores e não como verdades absolutas. Assim, de modo geral, o objetivo do trabalho, que era o de analisar a opinião dos usuários do *Twitter* sobre o Exame Nacional do Ensino Médio (ENEM) 2016 através da técnica de Mineração de Textos usando *Clusterização* para extrair conhecimento inerente aos textos analisados, foi atingido.

6.1. Contribuição

A partir dos resultados obtidos na utilização do modelo Cassiopeia como clusterizador de textos, pode-se observar algumas contribuições para a área de mineração de textos:

- Aplicação do Modelo Cassiopeia como opção para a clusterização de textos;
- O processo de coleta e estruturação dos *tweets* em Liguagem Natural, que é ambígua e apresenta uma série de informações desnecessárias, mostrando que os métodos de KDT são extremamente dependentes de técnicas de pré-processamento dos textos para padronizá-los e representá-los de forma mais concisa e fácil de ser interpretada pelos usuários;
- A sistematização do processo de mineração de texto através do Modelo Cassiopeia indicando quais as etapas e técnicas associadas a cada uma delas;

- A realização da análise qualitativa dos *clusters* através da nuvem de palavras;
- A descoberta de conhecimentos significativos nos *tweets* sobre o ENEM;
- Publicação do artigo: SILVA, L. M; GUELPELI, M. V. C. **COLETA NA REDE SOCIAL *TWITTER*: um *corpus* do *Enem 2016.*** VII Congresso Internacional de Conhecimento e Inovação 2017 – Foz do Iguaçu/PR.
- Além disso, o processo apresentado neste trabalho pode ser seguido por organizações para coletar a opinião de usuários do *Twitter*, fazendo o uso dos resultados para os mais diversos fins dentro das mesmas.

6.2. Dificuldades e Limitações

Reconhece-se como uma limitação deste trabalho a não utilização de um avaliador humano especialista no processo do ENEM. Houve dificuldade de obter uma co-orientação desse especialista, para acompanhar e analisar os resultados no pós-processamento. Acredita-se que os resultados poderiam ser melhores analisados utilizando mais métricas.

Outra limitação desta pesquisa é a utilização de apenas parte do *corpus* para o processo de clusterização.

6.3. Trabalhos Futuros

Com a dificuldade e limitação de se conseguir um especialista para avaliar os agrupamentos gerados no processamento do Modelo Cassiopeia, surgiu a possibilidade em trabalhos futuros, realizar a análise por um especialista para encontrar informações novas e implícitas dentro de cada *cluster*. Com isso, as métricas externas também seriam utilizadas.

Também pretende-se avaliar o processo de descoberta de conhecimento com o *corpus* inteiro de documentos textuais construído neste trabalho, já que a mineração de textos foi realizada em amostras de 700 *tweets* em cada categoria.

Outra possibilidade de trabalhos futuros que serão embasados no trabalho atual, integrar várias bases de dados, como a do Censo Escolar, ENEM e do Censo de Educação Superior, a fim de identificar padrões e modelos que suportem a criação de estratégias para a melhoria do ENEM, bem como auxiliar, por exemplo, os gestores públicos na criação/consolidação de ações afirmativas sobre o exame.

REFERÊNCIAS

ALUÍSIO, Sandra M.; ALMEIDA, Gladis M.B**. O que é e como se constrói um *corpus*? Lições aprendidas na compilação de vários corpora para pesquisa linguística.** Calidoscópio, v. 4, 2006. Disponível em: <http://revistas.unisinos.br/index.php/calidoscopio/article/view/6002> Acesso em: 12 de Mai. 2016.

ARANHA, Christian; PASSOS, Emmanuel. **A Tecnologia de Mineração de Textos.** Lab.ICA Elétrica PUC-Rio. RESI-Revista Elerônica de Sistemas de Informação, Nº2, 2006. Disponível em: <www.periodicosibepes.org.br/index.php/reinfo/article/download/171/66>. Acesso em: 17 de Jun. 2016

ARANHA, Christian Nunes. **Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português.** Tese (Doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, 2007. Disponível em: <livros01.livrosgratis.com.br/cp064589.pdf>. Acesso em: 12 de Jul. 2016.

BAUER, Martin W.; ARTS, Bas. **A construção do *corpus*: um princípio para a coleta de dados qualitativos.** Pesquisa qualitativa com texto, imagem e som. Petrópolis: Vozes, 2002. Disponível em: < http://discovery.ucl.ac.uk/60218/>. Acesso em: 25 de Jul. 2016.

CORRÊA, Adriana Cristina Giusti. **Recuperação de documentos baseados em informação semântica no ambiente AMMO**. São Carlos: UFSCar, 2003. Disponível em: <https://repositorio.ufscar.br/bitstream/handle/ufscar/522/DissACGC.pdf?sequence=1&isAllowed=y> Acesso em: 25 de Jul. 2016.

DIREITOS BRASIL. **Ocupação nas Escolas: O que representam? Quais os objetivos?** Disponível em: <http://direitosbrasil.com/ocupacao-nas-escolas/#forward> Acesso em 22 de Dez. 2016.

FACELI, Katti; LORENA, Ana C.; GAMA, João; CARVALHO, André C. P. L. F. **Inteligência Artificial: Uma Abordagem de Aprendizagem de Máquina**. Ed. LTC, 2011. 378 p. Vol. Único.

GOMES, Helder Joaquim Carvalheira. **Text Mining: análise de sentimentos na classificação de notícias.** Information Systems and Technologies (CISTI), 2013. Disponível em: <https://run.unl.pt/bitstream/10362/9182/1/TEGI0325.pdf>. Acesso em: 10 Set. 2016.

GOMIDE, Janaina Sant'Anna. Mineração de redes sociais para detecção e previsão de eventos reais. Belo Horizonte, 2012. Disponível em: < https://www.dcc.ufmg.br/pos/cursos/defesas/1491M.PDF >. Acesso em: 05 de Set. 2016.

GONÇALVES, Teresa. *et al*. **Analysing Part-Of-Speech for Portuguese Text Classification.** In: Computational Linguistics and Intelligent Text Processing. Springer, 2006. Disponível em: <https://www.researchgate.net/publication/221629265_Analysing_Part-of-Speech_for_Portuguese_Text_Classification>. Acesso em: 25 de Set. 2016.

GUELPELI, Marcus. V. C.; **Cassiopeia: Um modelo de agrupamento de textos baseado em sumarização.** Tese (doutorado) – Universidade Federal Fluminense. Programa de Pós-graduação em Computação, Niteroi-RJ, Brasil, 2012. Disponível em:< http://www.addlabs.uff.br/Novo_Site_ADDLabs/images/documentos/publicacoes/teses_dissertacoes/Marcus_guelpeli.pdf>. Acesso em: 29 de jul. 2016.

GUELPELI, Marcus. V. C.; FERNANDES, Heider. M. **Input a Word, Analyze the World: Selected Approaches to *Corpus* Linguistics.** 1ª. ed. United Kingdom: Cambridge Scholars Publishing, 2016. v. I. 521p.

HALKIDI, M.; BATISTAKIS Y.; VARZIRGIANNIS, M. **On *Cluster*ing validation techniques**. Journal of Intelligent Information Systems, 2001. Disponível em: < https://link.springer.com/article/10.1023/A:1012801612483>. Acesso em: 23 de Set. 2016.

LIDDY, E. *Natural Language Processing*. **Encyclopedia of Library and Information Science.** New York: Marcel Decker, Inc, 2001. Disponível em: < http://surface.syr.edu/cnlp/11/>. Acesso em: 15 de Nov. 2016.

LIMA, Jéssica R. C; MOURA, Karla H. S. **Mineração de Dados em Redes Sociais Usando o NODEXL.** Faculdade de Ciências e Tecnologia Mater Christi – Mossoró – RN – Brasil. 2014. Disponível em: <http://docplayer.com.br/8863498-Mineracao-de-dados-em-redes-sociais-usando-o-nodexl-data-mining-in-social-networks-using-nodexl-resumo.html.> Acesso em: 10 de Dez. 2016.

LUCHESI, André P; BERTOLA, José Renan; ARAÚJO, Juliano Leite. **Mineração de Textos na Web (Text Mining).** Pontifícia Universidade Católica de Campinas. 2006.

MANNING, C.D.; RAGHAVAN, P.; SHUTZE, H.; **Introduction to Information Retrieval**. Cambridge University Press, 2008. Disponível em: < http://www.math.unipd.it/~aiolli/corsi/0910/IR/irbookprint.pdf>. Acesso em: 12 de Dez. 2016.

MORAIS, Edison Andrade Martins; AMBRÓSIO, Ana Paula L. Mineraçao de textos. **Relatório Técnico–Instituto de Informática (UFG)**, 2007. Disponível em: < http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_005-07.pdf>. Acesso em: 15 de Jan. 2017.

PARANHOS, RANULFO et al. Uma introdução aos métodos mistos. **Sociologias**, v. 18, n. 42, 2016. Disponível em: < http://www.redalyc.org/html/868/86846760014/>. Acesso em: 10 Nov. 2017.

RODRIGUES, Paulo. R. F. **Dinamica de temas abordados no twitter via evolucão de clusters.** TCC (Graduação em Sistemas de Informação) - Universidade Federal do Ceará, 2016.

RUSSEL, Mathew A. ***Mining the social web*: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub and More.** 2 ed. Sebastopol: O'reilly Media, Inc., 2013. Disponível em: < Disponível em: <http://www.webpages.uidaho.edu/~stevel/504/Mining-the-Social-Web-2nd-Edition.pdf.>. Acesso em: 20 de Jan. 2017.

SANTOS, Wilian Pereira da Silva. Análise dos Tweets sobre a Black Friday através da Mineração de Texto e Análise de Sentimentos. Universidade Federal do Estado do Rio de Janeiro 2016. Disponível em: < http://bsi.uniriotec.br/tcc/textos/201601Wilian.pdf> Acesso em: 21 de Jul. 2017.

SARDINHA, Tony Berber. **Lingüística de *Corpus*: histórico e problemática.** *DELTA* [online]. 2000, vol.16, n.2. Disponível em: <http://dx.doi.org/10.1590/S0102-44502000000200005>. Acesso em: 21 Fev. 2017.

SAKAKI, T.; OKAZAKI, M. & MATSUO, Y. (2010). **Earthquake shakes *twitter* users: real - time event detection by social sensors**. Em Proceedings of the 19th international conference on World wide web, New York, NY, USA. Disponível em: < http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.183.8117&rep=rep1&type=pdf>. Acesso em: 10 de Mar. 2017.

SINCLAIR, J. 2005. ***Corpus* and Text - Basic Principles.** *In*: M.WYNNE (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford, Oxbow Books, p. 1-16. Disponível em: <http://ahds.ac.uk/linguistic-corpora/>. Acesso em: 12 de Fev. 2017.

SILVA, L.M; GUELPELI, M. V. C. **Coleta na rede social *twitter*: um *corpus* do *Enem 2016.*** VII Congresso Internacional de Conhecimento e Inovação 2017 – Foz do Iguaçu/PR.

TUMASJAN, A; Sprenger, T. O.; Sandner, P. G. & Welpe, I. M. (2010). **Predicting elections with *twitter* : What 140 characters reveal about political sentiment.** Word Journal Of The International Linguistic Association. Disponível em: < https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1441>. Acesso em: 24 de Fev. 2017.

TRASK, R.L. 2004. ***Dicionário de Linguagem e Lingüística.*** São Paulo, Contexto, 364 p.

VIANNA, Heraldo. M. **Avaliações nacionais em larga escala: análises e propostas.** Revista Estudos em Avaliação Educacional , n. 27, p. 41-76, jan./jul. 2003. Disponível em: < http://www.dma.ufv.br/downloads/MAT%20207/2016-I/textos/Texto%20complementar%20sobre%20avaliacoes%20sitemicas%20-%20MAT%20207%20-%202016-I.pdf>. Acesso em: 13 de Mar. 2017.

WIVES, Leandro K.; e OLIVEIRA, José P. M. **Aplicando métodos de Descoberta de Conhecimento em Textos em documentos sobre a mortalidade pública**. 1999. Disponível em: < http://seer.ufrgs.br/cadernosdeinformatica/article/view/v1n1p25-28> Acesso em: 10 de Jun. 2017.

WIVES, Leandro K. **Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (*Cluster*ing) de documentos**. 2004. Tese de Doutorado. Universidade Federal do Rio Grande do Sul. Disponível em: <http://www.lume.ufrgs.br/handle/10183/4576> Acesso em: 23 de Jun. 2017.